REMARKS

The Office Action mailed December 4, 2003 has been carefully reviewed and the foregoing amendments are made in response thereto. In view of the amendments and the following remarks, Applicants respectfully request reconsideration and reexamination of this application and the timely allowance of the pending claim.

The specification has been amended herein to update references to patent applications that have since issued as US patents.

Claims 1, 2, and 7 have been amended to more clearly claim the invention. Claim 1 has been amended to claim a set of tag probes selected from SEQ ID NO: 1-2000. This amendment is supported, for example, by figure 2 where SEQ ID NO: 1 is identified as a "tag probe" and its complement is identified as a "tag". Claim 1 has been further amended to add the limitation that the set comprises at least 1000 sequences from SEQ ID NO: 1-2000. This amendment is supported by original claim 4. Claim 2 has been similarly amended to claim a set of at least 1000 "tags" which are the complements of SEQ ID NO; 1-2000. Claim 7 has been amended to be dependent on claim 1.

Claims 15-19 are newly added. Support for claim 15 may be found in originally filed claim 7. Support for claims 16-19 may be found in figure 2 and on page 8 last line and page 9, lines 1-4. Claims 3-6 and claims 8-14 have been canceled herein.

Rejection of Claims 1-12 Under 35 USC 101

Claims 1-12 have been rejected by the Examiner as lacking patentable utility. Applicants respectfully disagree. The specific set of tags and tag probes claimed have both a well established utility and a substantial and specific utility disclosed in the specification. The set of sequences are useful as tag probes and their complements are useful as tags that hybridize to the tag probes. The set of tag probes, particularly when attached to a solid support so that each tag probe sequence is in a feature of known location, is useful to detect sequences that have been labeled with the complementary tag. The tags may be used to label, for example, primers which may be used for a variety of purposes, such as Single Base Extension (SBE). The tagged primers can be labeled in an assay specific manner and detected by hybridization to the array of tag probes. This utility is specific, asserted in the specification and well established.

In regards to the Examiner's observation that the sequences lack 100% specificity to any particular organism in GenBank, this is accurate, but this is a feature of the utility of the sequences. The sequences were selected so that they are not complementary to sequences in the public database so they will not hybridize to sequences that occur naturally in the genome. (See specification page 8, lines 23-24.) What is claimed is a set of tags and tag probes. The sequences are selected so that they hybridize under similar conditions to a corresponding tag probe and so that tags in the set hybridize only to their corresponding tag probe and not to other tag probes in the set. This allows the tagged sequence to be specifically detected by hybridization to its corresponding tag probe. Any genomic sequence can be tagged and then detected on an array of tag probes. The array of tag probes can be used to detect any tagged sequences and a researcher therefore has the flexibility to use the same array design to detect virtually any sequence that can be tagged. The same array can be used for many different experiments. See Exhibit 1 "GeneChip GenFlex Tag Array" product insert.

SEQ ID NO: 1-2050 were selected from a list of random sequence 20 mer sequences. Members of the set were selected so that each member of the set hybridizes to its complementary tag probe (the perfect complement of the tag over the 20 nucleotide length of the probe) with closely matched melting temperatures. Members of the set were also selected so that they do not cross hybridize to another member of the set or to sequences in the public database.

At least one specific utility is asserted in the specification. In the specification on page 10, lines 15-22 one use of tagged primers and an array of tag probes for genotyping is disclosed as follows (emphasis added):

> One embodiment of the method involves three step: (1) amplification of the polymorphic locus, (2) primer extension of a sequence-tagged primer with distinct labels for different polynucleotides at the polymorphic locus, and (3) hybridization to a tag array. The amount of each distinct label can be determined at known positions of the tag array. Each tag represents a distinct polymorphic locus and each distinct label represents a distinct allelic form at the polymorphic locus. The method permits the simultaneous determination of a genotype at multiple loci, as well as the determination of allele frequencies in a population. Another embodiment employs just steps (2) and (3).

In this embodiment, SEQ ID NO: 1-2050 are present on an array as tag probes and the complements of SEQ ID NO: 1-2000 are used as the tag portion of the sequence tagged primers. For each polymorphic locus to be analyzed a sequence-tagged primer with a unique tag is used.

The sequence tagged primers also have a 3' region that hybridizes immediately upstream of the polymorphic locus. The sequence tagged primers are hybridized to the amplified fragments containing the polymorphic loci and they are extended by a single base corresponding to the polymorphic base. Each species of dNTP present in the reaction is differentially labeled. The labeled sequence-tagged primers are then hybridized to the array of tag probes and the hybridization pattern is analyzed. The tag probes are present at known locations so the genotype of each polymorphic allele may be determined by determining which label or labels have hybridized to the corresponding tag probe. This method was used in Fan et al. a copy of which has been provided with this response.

The utility of the claimed invention is also well established. Exhibit 3 includes a list of publications describing studies that have used the GenFlex tag array or sequences from SEQ ID NO: 1-2000. The prior art, for example, Shoemaker et al. *Nat. Genet.* 14:450-456 (1996), a copy of which has been provided with this response, demonstrates the well established utility of sets of 20 base tag sequences that can be detected by hybridization to a tag array and that can be used as unique identifiers.

Finally, the Examiner indicates that there is nothing specific to the 2050 sequences that distinguishes them from a different set of 2050 sequences or from any set of 2010 unrelated sequences. Applicants respectfully disagree. This set of 2050 sequences was carefully selected to match a selected set of criteria. They are all the same length, they were not present in the GenBank database, they have closely matched melting temperatures so that they are capable of specifically hybridizing to their corresponding target sequence under the same hybridization conditions, allowing the hybridization of all 2050 probes to be done in parallel in the same reaction, and no two sequences in the 2050 sequences will hybridize to the same target under the selected hybridization conditions. There is an almost infinite number of sequences that would not match each of these criteria and would therefore not be suitable for use as tag-tag probe sequences. SEQ ID NO: 1-2050 were further selected from a group of 2200 sequences that met the criteria described above, based on the ability of those 2050 to discriminate between the perfect match and mismatch probes. For additional details on how SEQ ID NO: 1-2050 were selected see Exhibit 2, GeneChip GenFlex Tag Array Technical Note No. 1.

Applicants respectfully request withdrawal of the rejection.

Rejection of Claims 1-12 Under 35 USC 112

Claims 1-12 have been rejected as not complying with the enablement requirement. The Examiner asserts that the claimed subject matter was not described in such a way as to enable one skilled in the art to make and or use the invention. Applicants respectfully disagree. The invention as claimed is a set of sequences that can be used as tag probes and their complements (tags). Each member of the set hybridizes specifically to its complement under similar conditions and does not cross hybridize with the complement of any other member of the set under the selected hybridization conditions. The members of the set also do not hybridize to sequences present in the public databases under stringent hybridization conditions. The set has also been empirically tested so that the members of the set are known to discriminate between the perfect match probe and the mismatch probe with high efficiency.

The specification provides at least two ways that the tags and tag probes may be used. First they may be used to tag individual deletion mutations. This has been used in yeast to identify open reading frames that are essential for growth under selected conditions. Briefly, each strain has a single deletion, each deletion is tagged with a different tag, the different strains are grown in the same culture in the selected conditions. The nucleic acid is isolated and analyzed using an array of tag probes. The tags that are present indicate deletions that are tolerated under the selected conditions and the tags that are absent indicate deletions that are not tolerated under the selected conditions. Second the tags may be used to label primers for locus specific genotyping as described above. Both of these uses are described on page 10 of the specification. Applicants respectfully request withdrawal of the rejection.

CONCLUSION

In view of the foregoing amendments and remarks, Applicants believe all pending claims are now in condition for allowance and should be passed to issue. If the Examiner feels that a telephonic interview would in any way expedite the prosecution and allowance of this application, please do not hesitate to call the undersigned at (408) 731-5768. The Commissioner is hereby authorized to charge any additional fees which may be required, or credit any overpayment to Deposit Account 01-0431.

Dated: March 4, 2004

Respectfully submitted,

Sandra E. Wells
Reg. No. 52,349

Affymetrix, Inc.
3380 Central Expressway
Santa Clara, CA 95051
Tel: 408-731-5000
Fax: 408-731-5392

EXHIBIT 1

GeneChip® GenFlex™ Tag Array

Flexible GeneChip® Array

The GenFlex™ Tag Array (P/N 900302) is the first in a powerful new line of flexible GeneChip® probe arrays from Affymetrix. The GenFlex™ Tag Array uses nucleic acid hybridization to parse as many as 2000 individual reaction products from complex pooled mixtures. You can now create and implement hybridization-based assays of your own design.

Unlimited Applications

The GenFlex™ Tag Array is virtually unlimited in its potential to support new applications. Any reaction in which a nucleic acid tag can be incorporated into the products may be amenable to analysis with the array. GeneChip® probe arrays are ideally suited for separating complex mixtures, enabling highly multiplexed assay strategies.

Examples of applications considered include the following:

- Analysis of mutant strains of *Saccharomyces cerevisiae*[1]
- Genotyping of single nucleotide polymorphisms (SNPs)[2]

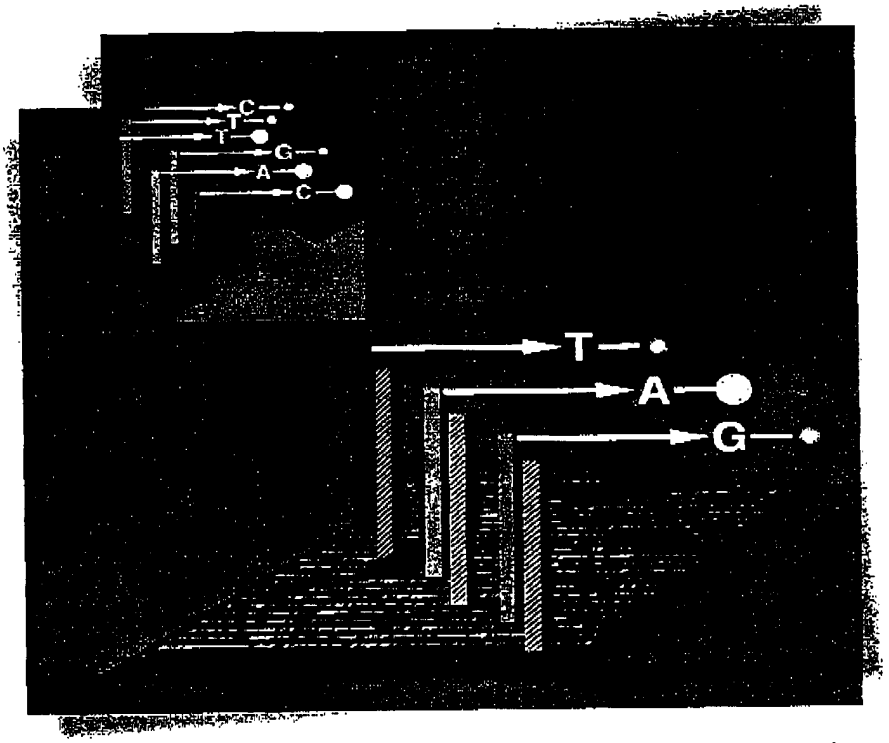


Figure 1. The GenFlex™ Tag Array Approach. A typical assay is depicted in which the Tag sequences (colored segments) have been incorporated into a set of reaction products. In this example, a set of nucleic acid sequences have been interrogated by a biochemical reaction resulting in the addition of a nucleotide (A,C,G, or T) to each Tag-containing molecule. The nucleotide is labeled with one of two different fluorophores (red and yellow circles). The assay may be carried out in multiplex reactions which are then combined into a single pool. The result of each individual reaction (i.e., which labeled base was incorporated) can be determined by hybridization to the GenFlex™ Tag Array. The Tag-probe (stippled colored segments) on the GenFlex™ Tag Array captures the corresponding reaction product by hybridizing to the Tag sequence. The hybridization signal detected for each of the 2000 Tags on the GenFlex™ Tag Array reveals the results of each individual reaction.

Affymetrix has partnered with Orchid BioSciences, Inc. to develop SNP-IT™ assays, analogous to the single base extension strategy employed by Fan *et al*, for SNP genotyping for use with the GenFlex™ Tag Array.

Note: *A license is required from Orchid BioSciences, Inc. to conduct single base extension assays, called SNP-IT™ assays, with GeneChip® GenFlex™ Tag Arrays.*

AFFYMETRIX®

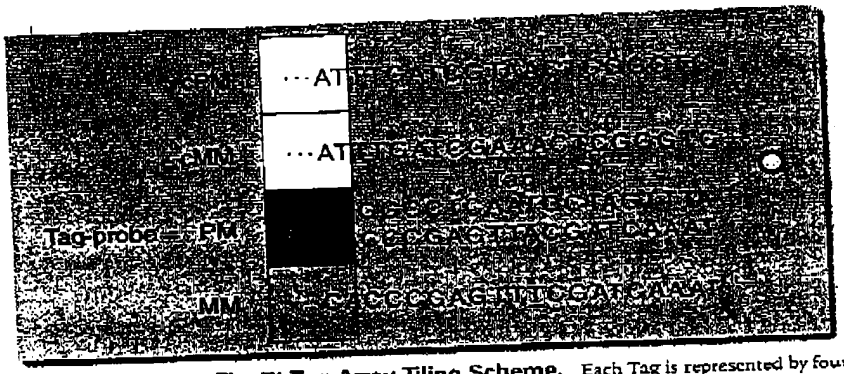


Figure 2. The GenFlex™ Tag Array Tiling Scheme. Each Tag is represented by four features arranged vertically on the GenFlex™ Tag Array. The Tag-probe designed to be the perfect complement to the Tag, is called the "PM" or perfect match probe in the Affymetrix® Microarray Suite software. The mismatch-containing Tag-probe with a different base at position 11 (from the 5' end as depicted here) is referred to as "MM" in the software. Two additional features are included on the array and reported in the software, the "CPM" and "CMM." These are the complements to the PM and MM tag-probes, respectively. The tag sequence, attached to a reaction product, hybridizes most strongly to the Tag-probe.

The Tag Array Strategy

The GenFlex™ Tag Array is comprised of capture probes for 2000 Tag sequences. These are 20mers which were selected from all possible 20mers to have similar hybridization characteristics and minimal homology to sequences in the public databases. The selection and characterization of the Tag sequences is further described in the Technical Note (P/N 700428) and to facilitate assay design the Technical Note and the Tag sequences are available on a CD-ROM (P/N 610026).

Any assay in which nucleic acid Tag sequences can be incorporated is a candidate application. 'Tag' refers to the sequence that is being captured by the array. The Tag-probe set refers to the four sequences on the array used to analyze a given Tag, one of which is the perfect complement of the Tag (figure 2).

Typically, the Tag sequence is covalently attached to a molecule of interest, for example a DNA primer. The linked molecule then is the target of a biochemical reaction in which a fluorophore is incorporated into the Tag-containing molecule. The reaction products, up to 2000, are pooled and hybridized to the GenFlex ™ Tag Array. The results of each individual reaction can then be discerned by the hybridization signal intensity detected for the corresponding Tag-probe sets.

Fifty fluorescein-labeled Tag oligonucleotides are included and should be used as hybridization controls with every array.

Benefits

- **Flexible** – Design your own assays with a standard array.
- **Highly parallel** – Up to 2000 reactions analyzed in a single experiment.
- **Cost-effective** – The array is competitively priced and supports cost-reducing multiplex strategies to enable effective high-throughput assays.
- **Proven platform** – The GeneChip® system represents the state of the art in accurate, reproducible microarray analysis.

Specifications

Feature size	50 micron
Array size	mini
Oligo length	20mers
Control sequences included	
Hybridization controls	GenFlex™ Tag Array Reagent Kit

Ordering Information

References

[1]Shoemaker *et al.* Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genetics* 14, 450-456 (1996).

[2]Fan *et al.* Parallel genotyping of human SNPs using generic oligonucleotide tag arrays. *Genome Research* 2000 June; 10(6): 853-60.

AFFYMETRIX, INC.
3380 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-362-2447 (1-888-DNA-CHIP)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd.,
Voyager, Mercury Park,
Wycombe Lane, Wooburn Green,
High Wycombe HP10 OHH
United Kingdom
Tel: +44 (0)1628 552550
Fax: +44 (0)1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

www.affymetrix.com

For research use only.
Not for use in diagnostic procedures.

Part No. 700427 Rev 2

AFFYMETRIX®

EXHIBIT 2

TECHNICAL NOTE No. 1

GeneChip®
GenFlex™ Tag Array

The GenFlex™ Tag Array (P/N 900302) enables the interrogation of up to 2000 nucleic acid reaction products and 50 control oligonucleotides. This document describes the selection and evaluation of the Tag-probe sequences included on the array. Here 'Tag' refers to the sequence that is being captured by the array, i.e., a sequence that is attached to a reaction product of interest. The 'Tag-probe' refers to the sequence on the array, the complement of the Tag. Thus, the reaction product into which the Tag has been incorporated will hybridize to the corresponding Tag-probe on the array. The Tag sequences to be used in primer design are provided free of charge on a CD-ROM (P/N 610026).

Tiling Strategy

Four different sequences are tiled on the array to interrogate each Tag (see figure 1, Tiling strategy). The Affymetrix® Microarray Suite software output, the Hybridization Analysis Window (HAW), reports the background subtracted signal intensities for each of these features. The probe designed to be the perfect match complement to the Tag, (i.e., Tag-probe) is referred to as "PM" in the HAW of Affymetrix® Microarray Suite software. The probe designed to be the mismatch-containing Tag-probe with a different base at position 10 is referred to as "MM" in the HAW. Two additional control probes are included on the array and reported in the HAW, the "CPM" and "CMM." These probes are designed to be the complements to the PM and MM Tag-probes, respectively.

Sequence Selection

An initial set of 20mer Tag-probe sequences was selected with closely matched melting temperatures. A further filter, based on Affymetrix expression rules and probe array experience, was applied to optimize and standardize the hybridization characteristics of the set. Finally, Tag-probe sequences were removed if they were identical or nearly identical to each other, to control sequences used on the array borders, or to sequences in the public databases at the time of array design. The GenFlex™ Tag Array was designed with 2200 20mer sequences.

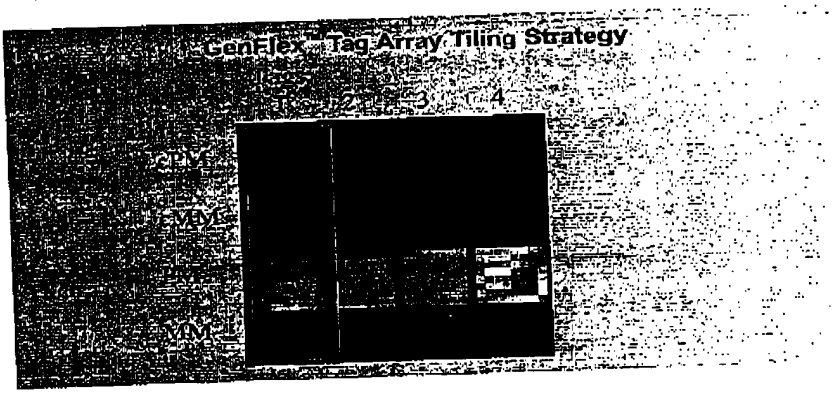


Figure 1. GenFlex™ Tag Array Tiling Strategy

Actual probe array features from a GenFlex™ Tag Array are shown. This array was hybridized with biotin-labeled oligonucleotides complementary to the Tag-probes, stained with streptavidin-phycoerythrin, and the data collected with the laser scanner. Four features, organized vertically on the probe array, represent each Tag sequence (as described in the text). For each of the four tags shown, arranged horizontally across the probe array, the brightest hybridization signal is seen with the "PM" feature. This feature is comprised of the Tag-probe oligonucleotide, the perfect complement to the Tag. Some hybridization is also seen to the MM feature (note Tag 4).

Tag Performance

The hybridization performance of the entire set of 2200 Tag sequences was evaluated. Biotinylated oligonucleotides complementary to the PM features, and thus called 'Tags', were synthesized and hybridized to the GenFlex™ Tag Array. As shown in figure 2, the normalized natural logarithm of the signal intensities obtained are distributed about a geometric mean of 0.8 with a standard deviation less than 0.1. As shown in figure 3, the Tag hybridizations yield high PM/MM ratios (a measure of the specificity of the Tag hybridization) with greater than 98% generating a PM/MM ratio greater than 3/1. Figure 4 shows a plot of the discrimination score and the signal intensity for all 2200 sequences. A line was fitted to select the 2050 sequences with the highest discrimination AND signal intensity (i.e., those above and to the right of the fitted line). Fifty individual "control" sequences were selected from the 2050 sequences to be representative of the population with respect to observed signal intensities and discrimination (see figure 2 & 3). Tag sequences with relatively low signals are over-represented in the 50 control sequences so as to increase information about the sensitivity of experiments at the lower limit of detection. Fifty fluorescein-labeled oligonucleotides or Tags, (i.e., complementary to the control Tag-probes on the array) are included with the GenFlex™ Tag Array Kit (P/N 900302) and should be used as internal standards with every hybridization experiment.

The Tag-probe sequences, in a 3'-5' orientation, are provided in the GeneChip® Hybridization Analysis Window (see figure 5) for data tracking purposes only.

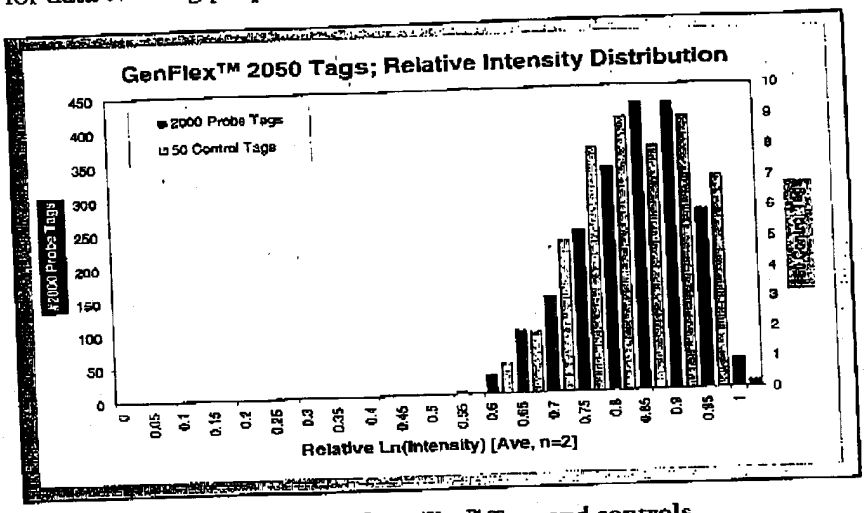


Figure 2. Signal intensities of GenFlex™ Tags and controls

Signal intensities were collected and averaged from two independent experiments in which the 2000 biotinylated Tag oligonucleotides or the 50 fluorescein-labeled control oligonucleotides were hybridized to the GenFlex™ Tag Array. The frequency of results are shown as normalized (to scale of 0-1, in bins of 0.05) natural logarithms of the net signal intensities.

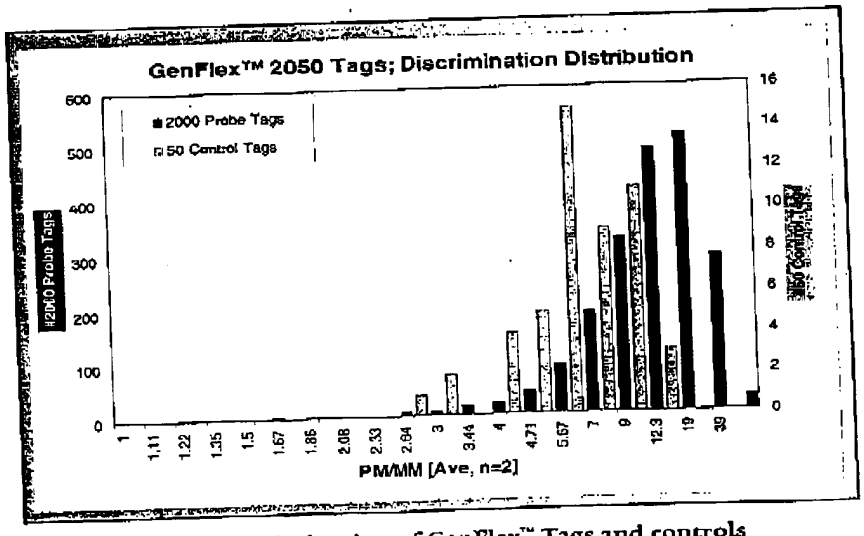


Figure 3. Signal discrimination of GenFlex™ Tags and controls

From the same data presented in figure 2, the PM/MM ratios are plotted. Greater than 98% of the Tags generate a ratio of greater than 3/1.

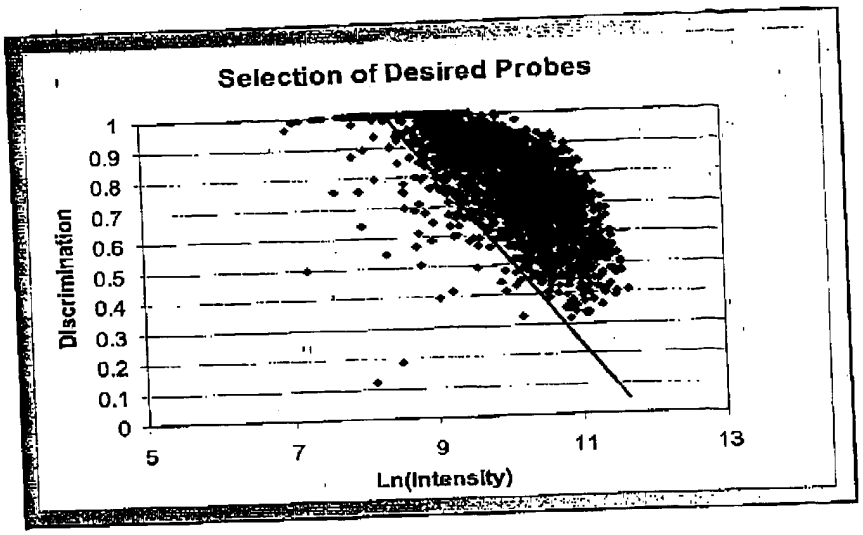


Figure 4. Selection of Tags

The natural logarithm of the average signal intensity and the discrimination score [(PM-MM)/(PM+MM)] is plotted for each of the 2200 Tags. A line was fitted to select the 2050 Tags with the highest signal intensities and highest discrimination scores.

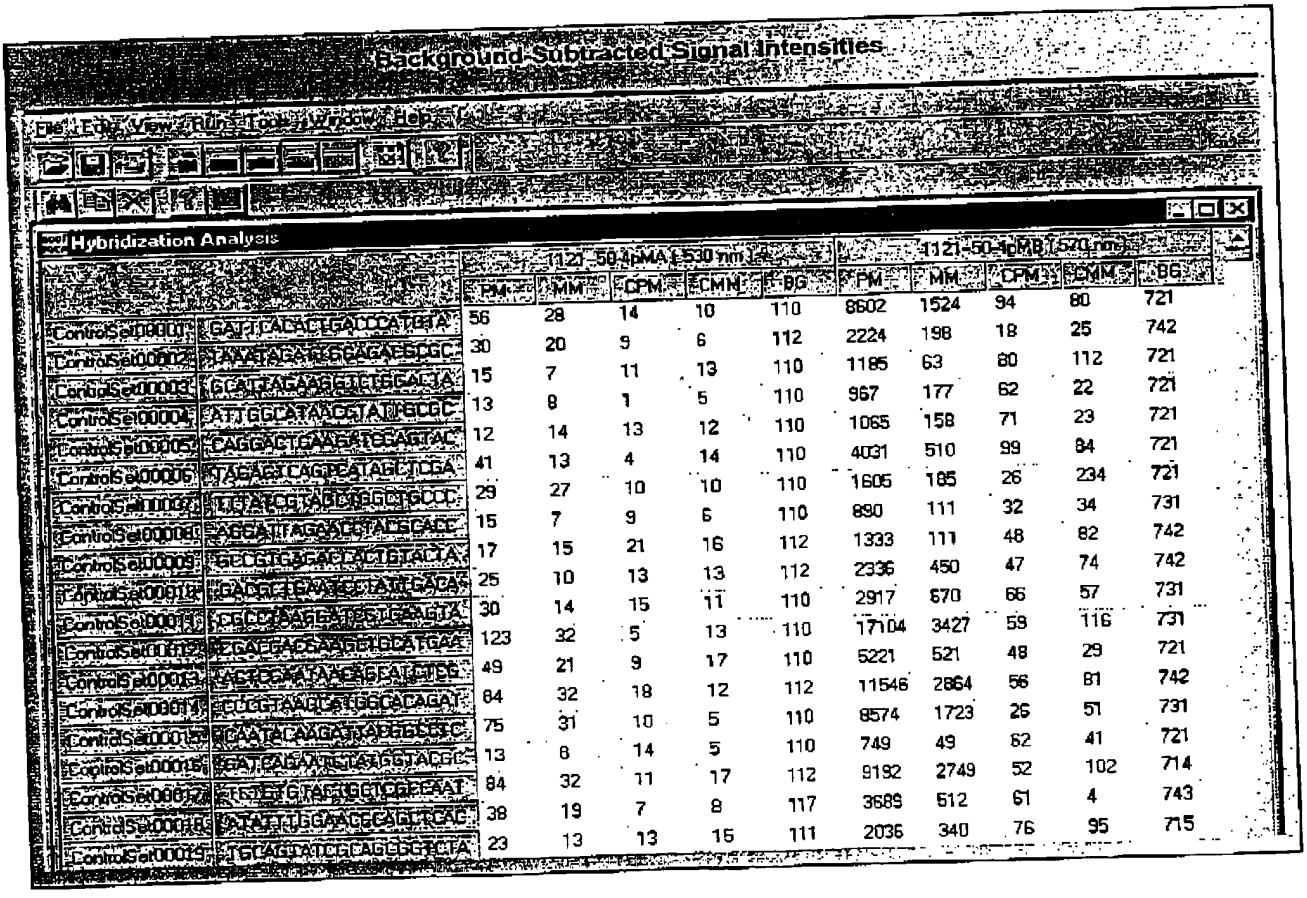


Figure 5. The Hybridization Analysis Window.

The background-subtracted signal intensities are shown for each of the four features (see text) 'PM, MM, CPM, CMM', for each Tag. 'BG' refers to the background signal intensity already subtracted from the 'PM, MM, CPM and CMM' values. Data are collected and reported for each of two different wavelength, 530 and 570 nm. The sequences shown in the HAW are those of the oligonucleotides synthesized on the probe array and are given in the 3' to 5' orientation.

Additional Information

GeneChip® probe array technology, products and services are marketed and sold internationally. For additional technical and pricing information about the HuSNP™ mapping assay and other GeneChip® probe array products, or to find the name and location of the distributor or sales office nearest you, please contact us at the following address:

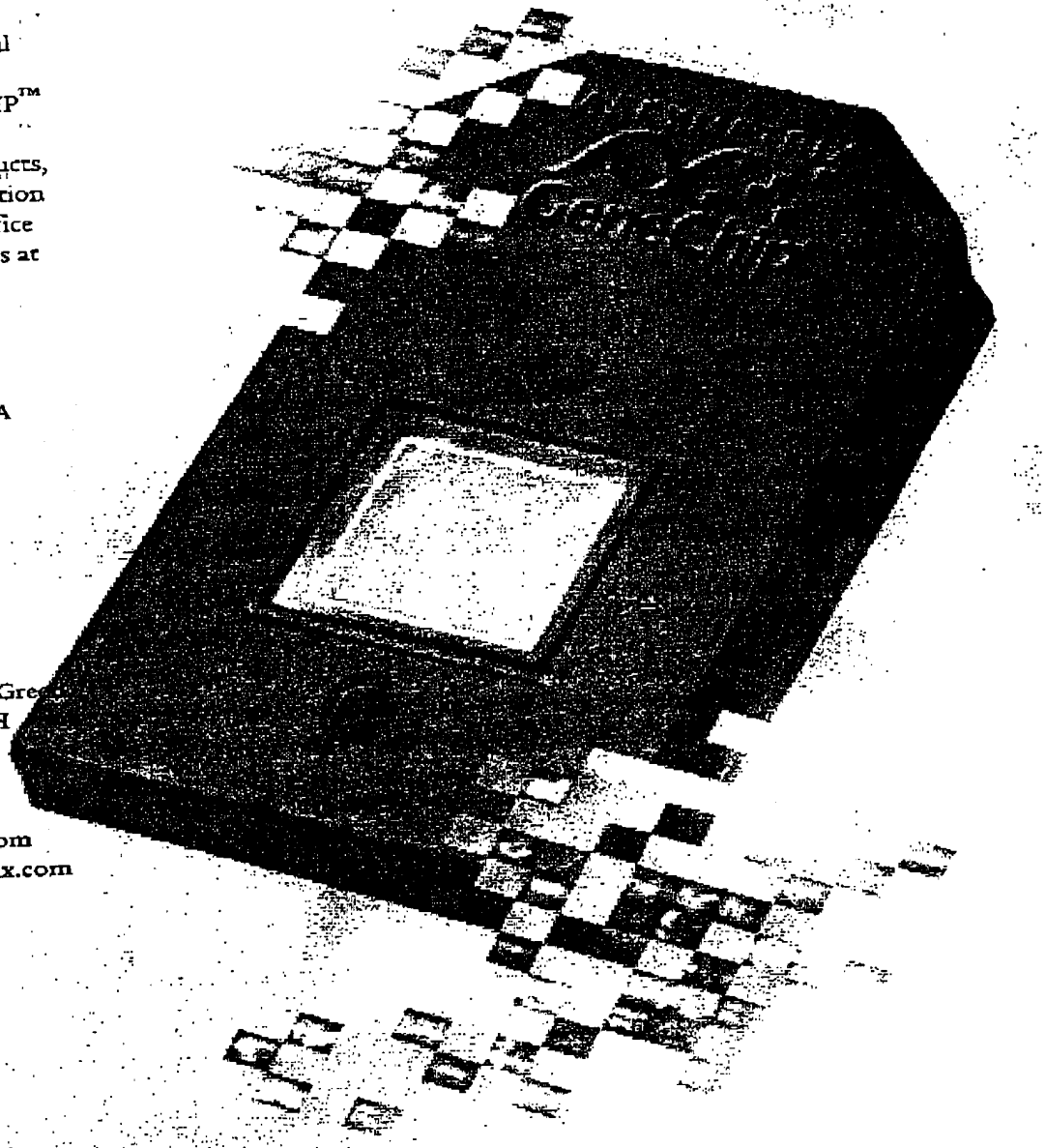
AFFYMETRIX, INC.

3380 Central Expressway
Santa Clara, CA 95051 USA
Tel: 1-888-362-2447
(1-888-DNA-CHIP)
Fax: 1-408-731-5441
sales@affymetrix.com
support@affymetrix.com

AFFYMETRIX UK Ltd.,

Voyager, Mercury Park,
Wycombe Lane, Wooburn Gre
High Wycombe HP10 0HH
United Kingdom
Tel: +44 (0)1628 552550
Fax: +44 (0)1628 552585
saleseurope@affymetrix.com
supporteurope@affymetrix.com

www.affymetrix.com



For research use only.
Not for use in diagnostic procedures.

AFFYMETRIX®

Exhibit 3 List of Publications Using GenFlex tag array

1. **Comparison of genflex tag array and pyrosequencing in SNP genotyping.**
Chen, D. C. *et al.* Journal of Molecular Diagnostics 5, 243-9, 2003.
2. **Parallel phenotypic analysis of sporulation and postgermination growth in Saccharomyces cerevisiae.**
Deutschbauer, A. M. *et al.* Proceedings of the National Academy of Sciences of the United States of America 99(24), 15530-5, 2002.
3. **Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays.**
Fan, J. B. *et al.* Genome Research 10(6), 853-60, 2000.
4. **Complementary whole-genome technologies reveal the cellular response to proteasome inhibition by PS-341.**
Fleming, J. A. *et al.* Proceedings of the National Academy of Sciences of the United States of America 99(3), 1461-6, 2002.
5. **Functional profiling of the Saccharomyces cerevisiae genome.**
Giaever, G. *et al.* Nature 418(6896), 387-91, 2002.
6. **Previously uncharacterized genes in the UV- and MMS-induced DNA damage response in yeast.**
Hanway, D. *et al.* Proceedings of the National Academy of Sciences of the United States of America 99(16), 10605-10, 2002.
7. **Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis.**
Ishii, M. *et al.* Genomics 68(2), 136-43, 2000.
8. **Multiplex SNP genotyping in pooled DNA samples by a four-colour microarray system.**
Lindroos, K. *et al.* Nucleic Acids Research 30(14), E70, 2002.
9. **Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots.**
Phillips, M. S. *et al.* Nature Genetics 33(3), 382-7, 2003.
10. **Significance analysis of microarrays applied to the ionizing radiation response.**
Tusher, V. G. *et al.* Proceedings of the National Academy of Sciences of the United States of America 98(9), 5116-21, 2001.

Letter

Parallel Genotyping of Human SNPs Using Generic High-density Oligonucleotide Tag Arrays

Jian-Bing Fan,[1,3] Xiaoqiong Chen,[1] Marc K. Halushka,[2] Anthony Berno,[1] Xiaohua Huang,[1,4] Thomas Ryder,[1] Robert J. Lipshutz,[1,6] David J. Lockhart,[1,5] and Aravinda Chakravarti[2]

[1]*Affymetrix, Inc., California 95051 USA;* [2]*Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106 USA*

Large scale human genetic studies require technologies for generating millions of genotypes with relative ease but also at a reasonable cost and with high accuracy. We describe a highly parallel method for genotyping single nucleotide polymorphisms (SNPs), using generic high-density oligonucleotide arrays that contain thousands of preselected 20-mer oligonucleotide tags. First, marker-specific primers are used in PCR amplifications of genomic regions containing SNPs. Second, the amplification products are used as templates in single base extension (SBE) reactions using chimeric primers with 3′ complementarity to the specific SNP loci and 5′ complementarity to specific probes, or tags, synthesized on the array. The SBE primers, terminating one base before the polymorphic site, are extended in the presence of labeled dideoxy NTPs, using a different label for each of the two SNP alleles, and hybridized to the tag array. Third, genotypes are deduced from the fluorescence intensity ratio of the two colors. This approach takes advantage of multiplexed sample preparation, hybridization, and analysis at each stage. We illustrate and test this method by genotyping 44 individuals for 142 human SNPs identified previously in 62 candidate hypertension genes. Because the hybridization results are quantitative, this method can also be used for allele-frequency estimation in pooled DNA samples.

The Human Genome Project and other private efforts are producing large amounts of genome sequence and polymorphism data that will provide scientists with an unprecedented opportunity to probe the structure and function of the human genome (Collins et al. 1998). In the realm of human disease, these genomic resources will allow the dissection of the genetic components and molecular mechanisms of complex human diseases and traits. Identification of complex disease genes will require both linkage and association analyses of thousands of polymorphisms across the human genome in thousands of individuals (Risch and Merikangas 1996; Collins et al. 1997; Chakravarti 1999). To enable such large-scale polymorphism analysis in human studies, parallel and efficient genotyping methods are critically needed. The most common variant in the human genome is the single nucleotide polymorphism (SNP) (Wang et al. 1998; Cargill et al. 1999; Halushka et al. 1999). Homogenous and microarray-based minisequencing has been used to genotype SNPs in human populations (Syvanen et al. 1990; Kuppuswamy et al. 1991; Chen and Kwok 1997; Pastinen et al. 1997,1998; Syvanen 1998). We present a parallel genotyping method for SNPs, termed TAG–SBE, which analyzes allele-specific single base extension (SBE) reactions on standardized, generic high-density oligonucleotide probe arrays (Chee et al. 1996; Shoemaker et al. 1996; Wang et al. 1998; Lipshutz et al. 1999). In TAG–SBE, the array is independent of the specific markers genotyped and the assay can be customized for sets of markers through PCR and SBE primer selection. Because this genotyping method is generic, intrinsically parallel, and favors multiplexed reactions, TAG-SBE is well-suited for large-scale human genetic studies.

To design the tag arrays, all possible 20 mers (4^{20} or ~10^{12}) were subjected to a computational screen that favored a subset of sequences with similar GC content and thermodynamic properties, and eliminated sequences with possible secondary structure or sequence similarity to other tags (Shoemaker et al. 1996; Gaever et al. 1999; Winzeler et al. 1999). A set of 32,000 tags was selected, with all tags expected to have similar hybridization characteristics and minimal cross-hybridization under standard hybridization conditions. As a hybridization control, and to enable background and cross-hybridization subtraction, each tag probe (PM, perfect match) is paired with a second probe that is identical in sequence except for a single base difference at the central position (MM, mismatch). The high-density tag array used in this study consists of over 64,000 distinct probes, over 32,000 PM tag probes, and over 32,000 adjacent MM probes, each probe occupying an area of 30 × 30 μm.

Present addresses: [3]*Illumina, Inc., San Diego, California 92121 USA;* [4]*Kiva Genetics, Inc., Mountain View, California 94043 USA;* [5]*Genomics Institute of the Novartis Research Foundation (GNF), San Diego, California 92121 USA.*
[6]**Corresponding author.**
E-MAIL rob_lipshutz@affymetrix.com; FAX (408) 481-0422.

10:853–860 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

The TAG–SBE genotyping method pairs the extension primer for each marker with a unique tag sequence, allowing the deconvolution of multiplexed preparations on a single high-density probe array (Fig. 1). The TAG–SBE approach can also be multiplexed both at the primary PCR and the SBE steps (see below). The resulting hybridization pattern from a typical TAG–SBE assay is shown in Figure 2A. The intensities of the two fluorophores used are measured and corrected for background and spectral overlap. The quantitative hybridization results are then used to make genotype calls (Figure 2B).

We first tested whether SBE methods for genotyping could be simplified. Previously published SBE methods such as minisequencing (Pastinen et al. 1997,1998; Syvanen 1998) and genetic bit analysis (Nikiforov et al. 1994; Head et al. 1997) required that double-stranded templates be converted to single-stranded templates prior to the base extension reaction [although double-stranded templates have been successfully used in fluorescence energy transfer-based SBE assays (Chen et al. 1997)]. We compared the TAG–SBE results obtained with three SNP markers using both single-stranded and double-stranded PCR products as templates, and found similar two-color intensity ratios and no significant differences in the absolute hybridization signal intensities. Thus, for all subsequent analyses, and the assays described here, double-stranded PCR templates were used in the SBE reactions.

To test the robustness, accuracy, and efficiency of the TAG–SBE method, we developed genotyping assays for a subset of the 874 SNPs that were identified recently in a large-scale polymorphism screen of 75 hypertension candidate genes (Halushka et al. 1999). Of these, we chose 171 SNPs in 68 genes, focusing on SNPs likely to have a functional significance: We chose SNPs in promoter regions, at splice junctions, and those that altered protein sequence. PCR primers were designed and tested individually for each of the 171 SNP-containing genomic regions. Of these, eight (4.7%) failed to amplify, and SBE primers were designed for the remaining 163 SNPs. We did not attempt to rescue the failed PCRs at this point, but this could be done if needed by reselecting primers or through a modification of the standard PCR conditions. For six of the 163 SNPs, SBE primers were designed for both the forward

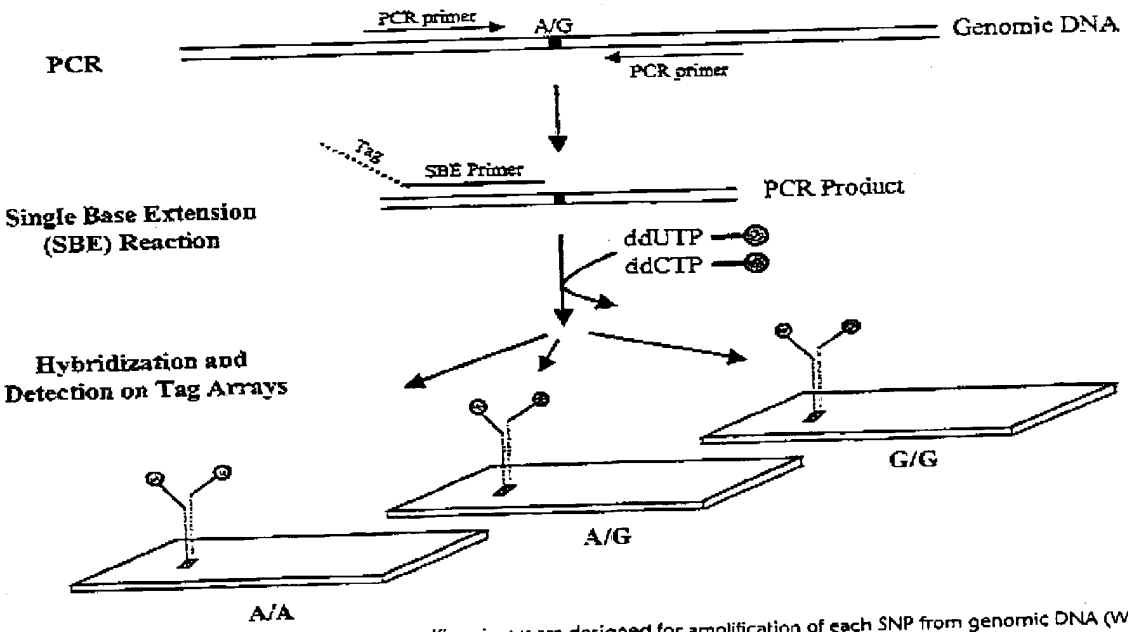


Figure 1 TAG–SBE genotyping assay. Marker-specific primers are designed for amplification of each SNP from genomic DNA (Wang et al. 1998); all SNPs with the same pair of variant bases (e.g., A/G SNPs) are pooled. The double-stranded PCR products serve as templates for the SBE reaction. Each SBE primer is chimeric with a 5′ end complementary to a unique tag synthesized on the array and a 3′ end complementary to the genomic sequence and terminating one base before a polymorphic SNP site. Thus, each SBE primer is uniquely associated with a specific tag (location) on the array. SBE primers corresponding to multiple markers are added to a single reaction tube and extended in the presence of pairs of ddNTPs labeled with different fluorophores; for example, an A/G bi-allelic marker is extended in the presence of biotin-labeled ddUTP and fluorescein-labeled ddCTP. The labeled multiplex SBE reaction products are pooled and hybridized to the tag array. Three hybridization patterns are shown, corresponding to three genotypes AA, AG, and GG.

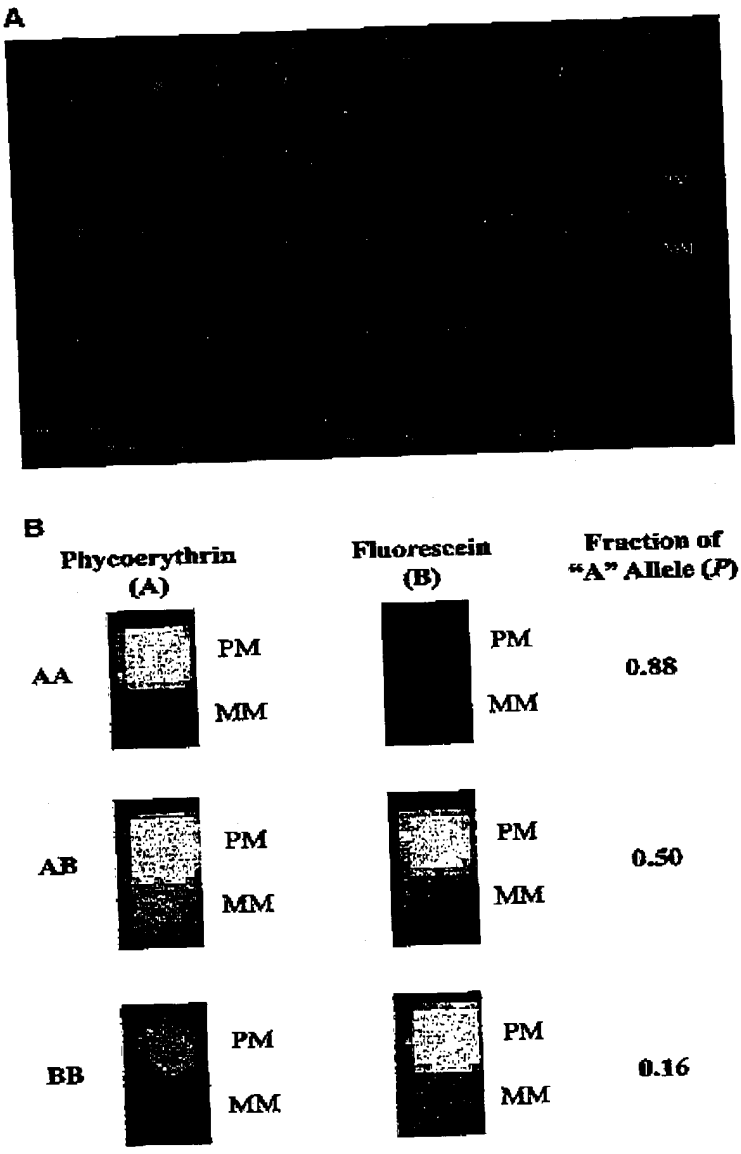


Figure 2 (*A*) Fluorescence image of a small portion of an oligonucleotide tag array following hybridization of 77 labeled SBE primers. The entire array contains >32,000 20-mer tag probe pairs. The physically adjacent PM and MM probes for a single tag probe pair are labeled. (*B*) The fluorescence intensity pattern for a tag probe pair showing the presence of an AA homozygote, an AB heterozygote, and a BB homozygote, and the computed relative allele fraction value $P = (PM - MM)_{fluorescein} / [(PM - MM)_{fluorescein} + (PM - MM)_{phycoerythrin}]$. Because of the partial overlap of the emission spectra of fluorescein and phycoerythrin, there is some spillover of fluorescein signal into the phycoerythrin emission channel. Background signals are subtracted and corrections for spectral overlap are applied prior to the quantitative genotyping analysis.

and reverse strands. Nine multiplex PCR and SBE reactions were designed with 9–28 markers in each set. Of the 163 SNP markers tested, 21 SNPs (12.9%) were further eliminated because they consistently produced poor signals in multiple samples tested. These failures were systematic, and were the result of poor amplification in the multiplex PCR or SBE reactions, or poor hybridization behavior on the array. It has been shown previously that roughly one out of 10 tag sequences do not hybridize sufficiently well on arrays of this type (Winzeler et al. 1999). Although these SNPs may be rescued by primer or protocol changes, repooling, using the opposite strand extension primer, or simply linking the primer to a different tag sequence (from which there are many to choose), we have not attempted further optimization of these 29 (8+21) markers. The remaining 142 markers in 62 genes were used in subsequent genotyping experiments. The 142 SNPs used, the genes involved and other details of the polymorphisms, and the designed primers are listed in a table located in the online supplement (note that the first 20 bases of the SBE primers listed in the table are complementary to the tag probes on the array). Additional information on these SNPs can be found indbSNP (http://www.ncbi.nlm.nih.gov/SNP/) or at http://genome.cwru.edu/candidates/snps.html3) (Halushka et al. 1999).

To test the reproducibility of the TAG–SBE assay, we performed the multiplex PCR, SBE reactions, and the array hybridization experiments in duplicate for four independent samples. A high correlation between the hybridization signals of the replicate measurements ($R^2 = 0.92$ for fluorescein signals and $R^2 = 0.93$ for phycoerythrin signals) was observed for the 142 SNPs. More importantly, there were no discrepancies in genotyping calls between the duplicate measurements.

We next used tag arrays to obtain the genotypes for all 142 SNPs in 44 unique DNA samples. Hybridization signals sufficiently above background were obtained for 96.5% (6029/6248) of the 6248 (142×44) possible calls. Based on the two-color signal intensity ratios, distinct genotype clusters were obtained for ~80% of the markers (Fig. 3). We used a combination of automatic software analysis and blind manual editing to assign genotypes for all 142 markers in the 44

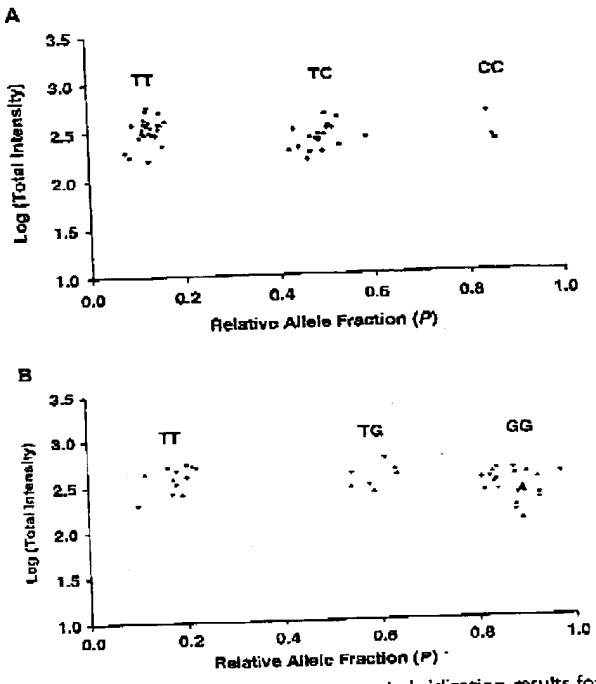


Figure 3 Cluster analysis of tag array hybridization results for 44 individuals at SNP marker (*A*) ANPex3.33 and (*B*) SELP.25. The logarithm of total fluorescence intensity [$(PM-MM)_{fluorescein} + (PM-MM)_{phycoerythrin}$] for each of the 44 hybridizations is plotted against the calculated relative allele fraction value *P*. The three distinct clusters observed correspond to the genotypes T/T, T/C, and C/C for marker ANPex3.33, and T/T, T/G, and G/G for marker SELP.25.

samples. For five of the six SNPs that had both forward and reverse SBE primers, identical genotypes were obtained from both strands in all 44 individuals (i.e., complete concordance in 220 paired tests). For one SNP (DCP1EX13.138), clear hybridization results were obtained for the forward primer, but the results were inconclusive for the reverse SBE primer and therefore calls for that strand were not made (i.e., one strand yielded clear results while the other produced a "no call"). In no cases did the two strands give contradictory results. This experiment indicates that either strand (or both) can be used for TAG–SBE analysis of the majority of the markers, and that for some markers, one strand may be more informative than the other. As described above, these assays were not fully optimized and we anticipate that it is possible to increase the overall genotyping yield further.

To determine the accuracy of the method, we used gel-based DNA sequencing to determine the genotypes of three individuals (a subset of the 44 persons studied earlier) at 133 loci. Comparison of the 355 paired gel-based and TAG–SBE genotype calls showed a total of 17 discrepancies involving seven different markers (see Table 1), a 4.8% discordance rate. Some of these discrepancies involved cases where one method made a homozygote call while the other method called a heterozygote. But there were also cases in which the gel-based sequencing and array-based genotyping yielded opposite homozygote genotype calls; we suspect systematic mispriming of the SBE primer to adjacent similar sequences as the likely cause of the discrepancy. Designing an SBE assay using primers for the other strand may be sufficient to solve the problem in most cases.

The quantitative nature of the two-color TAG–SBE measurements suggests the possibility of using pooled DNA samples to estimate allele frequencies and screen large numbers of loci for allele frequency differences between groups of phenotypically distinct individuals (Shaw et al. 1998 for microsatellite markers; Syvanen et al. 1993; Hacia et al. 1998 for SNP markers). To test this, we first synthesized two artificial SBE templates and performed controlled mixing experiments. As shown in Figure 4, the intensity ratio of the two fluorophores and the template concentration ratio are highly correlated over a 100-fold concentration range. We further tested the TAG–SBE assay performance with pooled DNA samples. Genomic DNA from five, 10, and 20 individuals with known genotypes was pooled and treated the same way as the DNA samples from individuals in all subsequent PCR amplification, SBE reac-

Table 1. Discrepancies Between Genotyping Calls with Gel-based Sequencing and the Array-based Method

			Gel-based sequencing			Array-based assay			
SNP name	WT allele	Mutant allele	904889	90896	904957	904889	904896	904957	Discrepancies
ACEEX17.19	C	A	C/C	A/A	A/C	C/C	C/C	C/C	2
CYP11B2EX6.91	T	C	T/C	T/C	T/C	T/T	T/T	T/T	3
CYP11B2BX7.65	T	C	T/C	C/C	C/C	T/T	T/T	T/T	3
GLUT4EX3.112	C	G	G/G	G/G	G/G	C/C	C/C	C/C	3
GALNREX1.553	G	C	G/C	G/G	G/C	G/G	G/G	G/G	2
ICAM1EX6.254	G	A	G/G	G/G	G/G	A/A	A/A	A/A	3
GMP-140.25	T	G	G/G	G/G	G/G	G/G	T/G	G/G	1

tion, and chip hybridization steps. In general, the observed allele frequencies were related directly to the values expected based on the known genotypes of the individuals in the pool (Fig. 5), and relatively small differences in allele frequency could be reliably detected for many markers. This strategy may be used to estimate allele frequencies in populations and to scan large numbers of markers for allele-frequency differences while greatly reducing the number of individual measurements required for association studies designed to detect genetic differences between groups of individuals with phenotypic differences. The minimum detectable allele-frequency differences and the maximum number of markers that can be genotyped in parallel remain to be determined.

Our approach combines the parallelism and flexibility of a standardized high-density oligonucleotide array readout with the enhanced fidelity of enzymatic primer extension reactions. Using a standard array of generic tags eliminates the need to design and manufacture custom arrays for specific sets of markers, as only the PCR and extension primers need to be customized. Furthermore, the tag-based approach uses as few as one or two oligonucleotide probes per marker rather than the 56 probes used previously on variant detector arrays (VDAs) (Wang et al. 1998). The standard tag array could also be used in combination with other genotyping approaches including multiplex oligonucleotide ligation assays (OLA) (Delahunty et al. 1996; Tobe et al. 1996; Chen et al. 1998), invasive cleavage of oligonucleotide probe assays (Lyamichev et al. 1999), and allele-specific PCR methods (Newton et al. 1989; Lo et al. 1991).

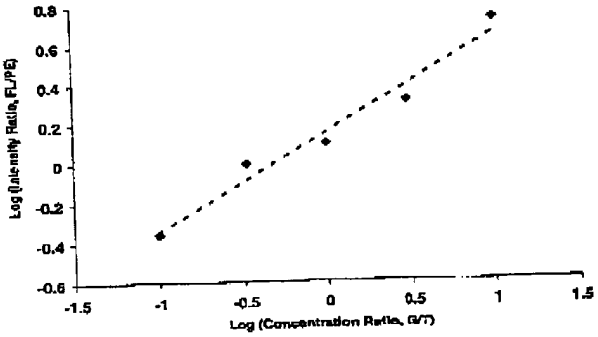


Figure 4 Quantitative allele frequency estimation based on two-color analysis of synthetic mixed templates. The two templates were mixed in the ratios of 1 nM/10 nM, 1 nM/3 nM, 1 nM/1 nM, 3 nM/1 nM, and 10 nM/1 nM, respectively. The logarithm of intensity ratios of the two colors (Y-axis) are plotted against the logarithm of concentration ratios of the two mixed templates (X-axis). FL, fluorescein intensity; PE, phycoerythrin intensity; G/T, concentration ratio of template G to template T.

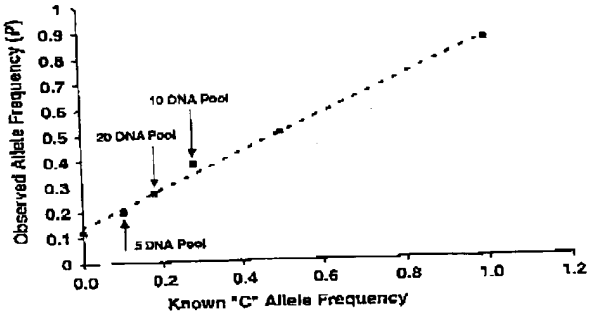


Figure 5 Allele frequency estimation for individual homozygotes, heterozygotes, and collections of multiple individuals at the SNP marker ANPex3.33. For the pooled samples, genomic DNA from a group of 5, 10, and 20 individuals (C-allele frequencies of 0.10, 0.28, and 0.18, respectively) was pooled in equal amounts and treated in the same way as the samples from single individuals. The observed allele fraction value P is plotted against the known C allele frequency, along with the best fit line as a guide to the eye. The line intercepts the Y-axis above the origin, and this systematic offset is the result of a small amount of cross-hybridization and misincorporation of the wrong base in the two-color SBE reaction. A correction can be applied to the data following the observation of pure genotypes to obtain a more accurate estimate of the absolute allele frequencies.

The experiments described here used only a small fraction of the 32,000 tags synthesized on the array and have not taken full advantage of the multiplexing possibilities. Our previous experience with developing highly discriminating sets of oligonucelotide probes for yeast gene expression measurements and genotyping, suggests that it should be possible to use a large fraction of the 32,000 tags on the array in a single experiment (Wodicka et al. 1997; Winzeler et al. 1998). A set of three such arrays would allow the determination of nearly 100,000 genotypes. The current array was synthesized using 30 μm features on an 8×8 mm chip. A single, 12.8×12.8 mm array with 24 μm features could interrogate 128,000 SNPs at a time. Physically smaller arrays with fewer tags may also be useful. Scaling down the array size to 2×2 mm, an array containing 24 μm features could encode over 3000 tags and accommodate many important genotyping applications in which more markers may not be necessary. In addition, multiple sets of tags can be associated with each locus-specific extension primer in separate reactions (pooled for hybridization). In this manner, a single array could be used to analyze the same loci from multiple individuals at once.

The highly parallel nature of oligonucleotide arrays and their ability to interrogate complex mixtures of nucleic acids enables significant flexibility in the design of genotyping assays. Simple calculations suggest that the cost of amplification and labeling reactions can be a significant barrier to the broad use of

large-scale genotyping methods. The multiplex sample preparations demonstrated here permit significant reductions in reagent use. Thus, multiplexing both specific genomic amplifications and SBE reactions reduced the 284 reactions needed for the 142 SNPs to only 18 reactions. This 16-fold reduction can be extended by pooling strategies. The current scheme uses two colors and requires six separate SBE reactions. The use of four colors would allow a single-tube reaction, with associated increases in efficiency and reduction of genotyping costs.

METHODS

Sample Collection and DNA Isolation

DNA samples from 44 individuals were collected as part of the ongoing GenNet network of the National Heart, Lung, and Blood Institute Family Blood Pressure Program. The sampling scheme was designed to ascertain nuclear families through a hypertensive proband. Samples were collected under informed consent and IRB approval at each of two field centers in Tecumseh, MI and Maywood, IL. DNA was extracted from buffy coats isolated from 5 to 10 ml of whole blood using a standard salting-out method and the PureGene kit (Gentra Systems). For the pooling experiments, genomic DNA from five, 10, and 20 individuals was pooled in equal amounts, and treated like single DNA samples in subsequent PCR amplifications, SBE reactions, and chip hybridizations.

Primer Design

For each SNP, primary PCR amplification primers were designed as described previously (Wang et al. 1998). The SBE primers were designed so that the 3′ end terminates one base before the polymorphic site. The Primer 3.0 software (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3.cgi) was modified and used to pick SBE primers at a predicted length of 20 nucleotides (range: 16–26) and melting temperature of 57°C (range: 53°C–64°C). SBE primers were picked from the forward direction first (i.e., 5′ to the SNP), the reverse direction being used when a suitable primer could not be chosen for the forward direction.

Multiplex PCR

Specific amplification of the genomic regions containing the 142 SNPs was achieved with nine multiplex PCR reactions, each containing 50 ng of human genomic DNA, 0.5 μM of each primer, 1 mM deoxynucleotide triphosphates (dNTPs), 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 5 mM $MgCl_2$ and 2 units of AmpliTaq Gold (Perkin Elmer) in a total volume of 25 μl. PCR was performed on a Thermo Cycler (MJ Research) with initial denaturation of the DNA templates and Taq enzyme activation at 96°C for 10 min, followed by 40 cycles of denaturation at 94°C for 30 sec, 57°C for 40 sec, and 72°C for 90 sec. The final extension reaction was at 72°C for 10 min.

SBE Template Preparation

One μl of Exonuclease I (10 U/μl, Amersham Life Science) and 1 μl of Shrimp Alkaline Phosphatase (1 U/μl, Amersham Life Science) were added to 25 μl PCR products and incubated at 37°C for 1 hr. The enzymes were inactivated at 100°C for 15 min. The enzymatically treated samples were applied to an S-300 column (Pharmacia) to further remove residual PCR primers and dNTPs. The buffer was replaced with ddH_2O.

Multiplex SBE Reaction

SBE reactions were carried out in 33 μl reactions using 6 μl of the template (see above), 1.5 nM of each SBE primer, 2.5 Units of Thermo Sequenase (Amersham), 52 mM Tris-HCl (pH 9.5), 6.5 mM $MgCl_2$, 25 μM of fluorescein-N6-d-dNTPs (New England Nuclear), 7.5 μM biotin-N6-d-dUTP or biotin-N6-d-CTP or 3.75 μM biotin-N6-d-dATP, and 10 μM of the other cold ddNTPs. Extension reactions were carried out on a Thermo Cycler (MJ Research) with 1 cycle at 96°C for 3 min, then 45 cycles of 94°C for 20 sec and 58°C for 11 sec. After SBE reactions, the products of the nine reactions from each sample were combined and mixed with 30 μl of 100 μg/ml glycogen (Boehringer Mannheim), 18.75 μl of 8 M LiCl (Sigma), and 1.1 ml of prechilled (−20°C) ethanol (200 proof), and precipitated by centrifugation (Eppendorf centrifuge 5415C) for 15 min at room temperature; precipitated samples were dried at 40°C for 40 min and resuspended in 33 μl ddH_2O.

Tag Array Design and Hybridization

For each tag sequence, two probes were synthesized on the array: one matches the designed-tag sequence exactly (PM probe) and the other being identical except for a single base difference in the central position (MM probe). The mismatch probe serves as an internal control for hybridization specificity and enables effective subtraction of background and cross-hybridization signals. Over 32,000 20-mer tag probes and their mismatch partners were chosen (Shoemaker et al. 1996) and fabricated on 8×8 mm arrays. Each probe (feature) occupies an area of 30×30 μm, which contains ~10^7 copies of the chosen 20-mer oligonucleotide. Sets of 100 arrays were synthesized together on a single glass wafer.

The labeled SBE reaction products were denatured at 95°C–100°C for 10 min and snap cooled on ice for 2–5 min. The tag array was prehybridized with 6× SSPE-T [0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA (pH 7.4), 0.005% Triton X-100] and 0.5 mg/ml BSA for a few minutes, then hybridized with 120 μl hybridization solution (shown below) at 42°C for 2 hr on a rotisserie (at 40 RPM). The hybridization solution consisted of 3M TMACl (tetramethylammonium chloride), 50 mM MES [2-(N-morpholinoethanesulfonic acid) sodium salt] (pH 6.7), 0.01% of Triton X-100, 0.1 mg/ml of herring sperm DNA, 50 pM of fluorescein-labeled control oligo, 0.5 mg/ml of BSA (Sigma) and 29.4 μl-labeled SBE products (see above) in a total value of 120 μl.

After hybridization, the arrays were rinsed twice with 1× SSPE-T for ~10 sec at room temperature, then washed with 1× SSPE-T for 15–20 minutes at 40°C on a rotisserie at 40 RPM. The arrays were washed 10 times with 6× SSPE-T at 22°C on a fluidics station (FS400, Affymetrix) and then stained at room temperature with 120 μl staining solution [2.2 μg/ml streptavidin R-phycoerythrin (Molecular Probes), and 0.5 mg/ml acetylated BSA, in 6× SSPET] and mixed on a rotisserie for 15 min at 40 RPM. After stainng, the arrays were washed 10 times with 6× SSPET on the fluidics station at 22°C. The arrays were scanned on a confocal scanner (Affymetrix) and fluorescence at 530 nm (fluorescein), and 560 nm (phycoerythrin) was collected with a spatial resolution of 60–70 pixels per feature. GeneChip software (Affymetrix) was used to convert image files into digitized files for further data analysis.

Genotype Determination

For a given marker (at a given tag probe position), the fluorescence intensity of each of the two fluorophores (fluorescein and phycoerythrin) was corrected for background and nonspecific hybridization by subtracting the intensity at the MM from that of the PM; negative values of PM–MM were treated as zero. Because of the overlap between the emission spectra of the two fluorophores, a fraction of the fluorescein signal (7.6%) was subtracted from the signal seen in the phycoerythrin channel (Hacia et al. 1998). A metric *P* which estimates the relative amount of each allele in the target mixture was computed as the relative proportion of the corrected intensities [fluorescein/(fluorescein+phycoerythrin)]. To define genotype clusters for each SNP (see Figure 3), the *P* values associated with each sample were sorted, and ranges corresponding to the three SNP genotypes were computed using an algorithm based on empirical observations across many genotyping experiments. The purpose of this algorithm is to identify well-separated ranges of experimental values that correspond to distinct genotypes. The specific algorithm employed here used the following rules: (1) At most four values (outliers), about 10% of the total data may be excluded from the computed ranges; (2) each pair of ranges must extend over an area of ≥0.3 and all three ranges must extend over ≥0.5; (3) individual ranges must be separated by a gap of ≥ 0.1; (4) the width of a single range may be ≤0.4. A "goodness" of fit statistic computed as 1 − (sum of range widths/total range) − (number of outliers /10) was maximized for the set of ranges chosen.

Quantitative Allele Analysis

Two templates, template-T (5′-TGCTGAATATTCAGATTCTCTAGTGCTACCTGAAAGATCCTG-3′) and template-G (5′-TGCTGAATATTCAGATTCTCGAGTGCTACCTGAAAGATCCTG-3′) were synthesized. They were identical except at a single (21st) position: T in template-T, and G in template-G. The two templates were mixed in the ratios of 1 nM/10 nM, 1 nM/3 nM, 1 nM/1 nM, 3 nM/1 nM, and 10 nM/1 nM, respectively. The following five distinct SBE primers, 5′-TGCGATTCTTTGCCGTCAGGCAGGATCTTTCAGGTAGCACT-3′, 5′-GGCGAAGTTCCTCTAGTGTTCAGGATCTTTCAGGTAGCACT-3′, 5′-GGCCTCGGTGTTCAGCATATCAGGATCTTTCAGGTAGCACT-3′, 5′-TGGAGATCGTTGCTTGTACCCAGGATCTTTCAGGTAGCACT-3′, 5′-TGCATTGATTTAACTGCGCGCAGGATCTTTCAGGTAGCACT-3′, were added separately to five SBE reactions containing the five types of mixed templates. The SBE primers were extended in the presence of biotin-labeled ddATP and fluorescein-labeled ddCTP, pooled, and hybridized to a tag array.

Gel-based Automated DNA Sequencing

To independently confirm the genotypes called using the TAG–SBE assay, three samples (904957, 904896, and 904889) were sequenced for 115 SNPs from the table in the online supplement, using conventional gel-based methods. Samples were amplified for all sites with T7- and T3-tagged primers using standard PCR cycling conditions [2.5 µl of 20 ng/µl DNA, 0.375 µl of 20 µM primer (X2), 1.5 µl of 10× PCR buffer, 0.9 µl 25mM $MgCl_2$, 0.15 µl 10 mM dNTPs, 0.25 µl 10 U/µl Taq DNA Polymerase (Sigma), in a total volume of 15 µl with ddH_20]. Some products were sequenced directly while others required an M13 nesting strategy because of the close proximity of the polymorphic base and primer end. Samples from the initial amplification were diluted 1:50 with ddH_20 and amplified with M13F-T7 (5′-TGTAAAACGACGGCCAGTTAATACGACTCACTATAGGGAGA-3′) and M13R-T3 (5′-AACAGCTATGACCATGAATTAACCCTCACTAAAGGGAGA-3′) primers using standard PCR conditions. All PCR products were cleaned with Exonuclease I (Amersham 0.15 µl of 10 U/µl per well) and Shrimp Alkaline Phosphatase (Amersham, 0.30 µl of 1 U/µl per well) in a volume of 10 µl. Dye terminator sequencing using an M13R primer (AACAGCTATGACCATG) or T7 primer (TAATACGACTCACTATAGGGAGA) on an ABI377 (Perkin Elmer) using Big Dye (Perkin Elmer) was performed to determine the genotype status for each SNP in each of the three individuals. Trace files were read with Edit View 1.0 (Perkin Elmer) software.

ACKNOWLEDGMENTS

We thank Drs. A. Weder and R. Cooper for DNA sample collection, M. Mittmann and D. Shoemaker for tag selection and array design, D. Stern for construction of array scanners used in this study, K. Bentley for DNA sequencing, and K. Gunderson for helpful discussions. This work was supported by a grant from the Advanced Technology Program of the National Institutes of Standards and Technology (70NANB5H1031) to Affymetrix, and research funds from Case Western Reserve University, University Hospitals of Cleveland, the National Heart, Lung, and Blood Institute (HL54466), and the National Institute of Mental Health (MH60007) to A.C. This research is a contribution of GenNet, a network of the National Heart, Lung, and Blood Institute's Family Blood Pressure Program.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22:** 231–238.

Chakravarti, A. 1999. Population genetics–making sense out of sequence. *Nature Genet.* **21:** 56–60.

Chee, M., R. Yang, E. Hubbell, A. Berno, X.C. Huang, D. Stern, J. Winkler, D.J. Lockhart, M.S. Morris, and S.P. Fodor. 1996. Accessing genetic information with high-density DNA arrays. *Science* **274:** 610–614.

Chen, X. and P.Y. Kwok. 1997. Template-directed dye-terminator incorporation (TDI) assay: a homogeneous DNA diagnostic method based on fluorescence resonance energy transfer. *Nucleic Acids Res.* **25:** 347–353.

Chen, X., B. Zehnbauer, A. Gnirke, and P.Y. Kwok. 1997. Fluorescence energy transfer detection as a homogeneous DNA diagnostic method. *Proc. Natl. Acad. Sci.* **94:** 10756–10761.

Chen, X., K.J. Livak, and P.Y. Kwok. 1998. A homogeneous, ligase-mediated DNA diagnostic test. *Genome Res.* **8:** 549–556.

Collins, F.S., M.S. Guyer, and A. Chakravarti. 1997. Variations on a theme: Cataloging human DNA sequence variation. *Science* **278:** 1580–1581.

Collins, F.S., A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters. 1998. New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282:** 682–689.

Delahunty, C., W. Ankener, Q. Deng, J. Eng, and D.A. Nickerson. 1996. Testing the feasibility of DNA typing for human identification by PCR and an oligonucleotide ligation assay. *Am. J. Hum. Genet.* **58:** 1239–1246.

Glaever, G., D.D. Shoemaker, T.W. Jones, H. Liang, E.A. Winzeler, A.

Astromoff, and R.W. Davis. 1999. Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nature Genet.* **21:** 278–283.

Hacia, J.G., K. Edgemon, B. Sun, D. Stern, S.P. Fodor, and F.S. Collins. 1998. Two color hybridization analysis using high density oligonucleotide arrays and energy transfer dyes. *Nucleic Acids Res.* **26:** 3865–3866.

Halushka, M., J-B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single nucleotide polymorphisms in candidate genes regulating blood pressure homeostasis. *Nature Genet.* **22:** 239–247.

Head, S.R., Y.H. Rogers, K. Parikh, G. Lan, S. Anderson, P. Goelet, and M.T. Boyce-Jacino. 1997. Nested genetic bit analysis (N-GBA) for mutation detection in the p53 tumor suppressor gene. *Nucleic Acids Res.* **25:** 5065–5071.

Kuppuswamy, M.N., J.W. Hoffmann, C.K. Kasper, S.G. Spitzer, S.L. Groce, and S.P. Bajaj. 1991. Single nucleotide primer extension to detect genetic diseases: experimental application to hemophilia B (factor IX) and cystic fibrosis genes. *Proc. Natl. Acad. Sci .* **88:** 1143–1147.

Lipshutz, R.J., S.P. Fodor, T.R. Gingeras, and D.J. Lockhart. 1999. High density synthetic oligonucleotide arrays. *Nature Genet.* **21:** 20–24.

Lo, Y.M., P. Patel, C.R. Newton, A.F. Markham, K.A. Fleming, and J.S. Wainscoat. 1991. Direct haplotype determination by double ARMS: Specificity, sensitivity and genetic applications. *Nucleic Acids Res.* **19:** 3561–3567.

Lyamichev, V., A.L. Mast, J.G. Hall, J.R. Prudent, M.W. Kaiser, T. Takova, R.W. Kwiatkowski, T.J. Sander, M. de Arruda, D.A. Arco et al. 1999. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nature Biotech.* **17:** 292–296.

Newton, C.R., A. Graham, L.E. Heptinstall, S.J. Powell, C. Summers, N. Kalsheker, J.C. Smith, and A.F. Markham. 1989. Analysis of any point mutation in DNA: The amplification refractory mutation system (ARMS). *Nucleic Acids Res.* **17:** 2503–2516.

Nikiforov, T.T., R.B. Rendle, P. Goelet, Y.H. Rogers, M.L. Kotewicz, S. Anderson, G.L. Trainor, and M.R. Knapp. 1994. Genetic bit analysis: a solid phase method for typing single nucleotide polymorphisms. *Nucleic Acids Res.* **22:** 4167–4175.

Pastinen, T., A. Kurg, A. Metspalu, L. Peltonen, and A.C. Syvanen. 1997. Minisequencing: A specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.* **7:** 606–614.

Pastinen, T., M. Perola, P. Niini, J. Terwilliger, V. Salomaa, E. Vartiainen, L. Peltonen, and A.C. Syvanen. 1998. Array-based multiplex analysis of candidate genes reveals two independent and additive genetic risk factors for myocardial infarction in the Finnish population. *Hum. Mol. Genet.* **7:** 1453–1462.

Risch, N. and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science* **273:** 1516–1517.

Shaw, S.H., M.M. Carrasquillo, C. Kashuk, E.G. Puffenberger, and A. Chakravarti. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.* **8:** 111–123.

Shoemaker, D.D., D.A. Lashkari, D. Morris, M. Mittmann, and R.W. Davis. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nature Genet.* **14:** 450–456.

Syvanen, A.C. 1998. Solid-phase minisequencing as a tool to detect DNA polymorphism. *Methods Mol. Biol.* **98:** 291–298.

Syvanen, A.C., K. Aalto-Setala, L. Harju, K. Kontula, and H.A. Soderlund. 1990. Primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. *Genomics* **8:** 684–692.

Syvanen, A.C., A. Sajantila, and M. Lukka. 1993. Identification of individuals by analysis of biallelic DNA markers, using PCR and solid-phase minisequencing. *Am. J. Hum. Genet.* **52:** 46–59.

Tobe, V.O., S.L. Taylor, and D.A. Nickerson. 1996. Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. *Nucleic Acids Res.* **24:** 3728–3732.

Wang, D.G., J-B. Fan, C.J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280:** 1077–1082.

Winzeler, E.A., D.R. Richards, A.R. Conway, A.L. Goldstein, S. Kalman, M.J. McCullough, J.H. McCusker, D.A. Stevens, L. Wodicka, D.J. Lockhart, and R.W. Davis. 1998. Direct allelic variation scanning of the yeast genome. *Science* **281:** 1194–1197.

Winzeler, E.A., D.D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J.D. Boeke, H. Bussey et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:** 901–906.

Wodicka, L., H. Dong, M. Mittmann, M.H. Ho, and D.J. Lockhart. 1997. Genome-wide expression monitoring in *Saccharomyces Cerevisiae*. *Nat. Biotechnology.* **15:** 1359–1367.

Received January 5, 2000; accepted in revised form March 29, 2000.

article

Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy

Daniel D. Shoemaker[1], Deval A. Lashkari[1], Don Morris[2], Mike Mittmann[2] & Ronald W. Davis[1]

A quantitative and highly parallel method for analysing deletion mutants has been developed to aid in determining the biological function of thousands of newly identified open reading frames (ORFs) in *Saccharomyces cerevisiae*. This approach uses a PCR targeting strategy to generate large numbers of deletion strains. Each deletion strain is labelled with a unique 20-base tag sequence that can be detected by hybridization to a high-density oligonucleotide array. The tags serve as unique identifiers (molecular bar codes) that allow analysis of large numbers of deletion strains simultaneously through selective growth conditions. Hybridization experiments show that the arrays are specific, sensitive and quantitative. A pilot study with 11 known yeast genes suggests that the method can be extended to include all of the ORFs in the yeast genome, allowing whole genome analysis with a single selective growth condition and a single hybridization.

The recent release of the complete nucleotide sequence of *Saccharomyces cerevisiae* provides researchers with the identities of all the genes in this basic eukaryotic cell. However, functional information is available for less than half of the estimated 6,000 genes[1]. Identifying the biological function of the thousands of genes discovered by the yeast (and other) genome sequencing projects is one of the main challenges facing researchers as we enter the post-genome era. Determining when a gene is expressed, where the gene product is localized in the cell, and what happens when the gene is disrupted are a few basic experimental questions that need to be addressed for the large numbers of newly discovered genes. There is a clear need for new genome-wide approaches that can provide answers to these questions in a rapid and cost effective manner.

Gene disruption is a powerful tool for determining the biological function of proteins encoded by uncharacterized ORFs. Given the sequence of a yeast ORF, it is possible to generate a precise null mutation (a deletion from start codon to stop codon)[2,3]. Information about the biological function can be inferred by monitoring the fitness of the resulting deletion strain under a variety of selective growth conditions (for example, exposure to UV light). This approach has been extremely effective for investigating the function of yeast ORFs. Unfortunately, the time and labour involved in analysing individual deletion strains will make it difficult to apply this approach on a genome-wide level.

Strategies have been reported in which gene disruptions[4] and selections are performed *en masse* using randomly integrated transposons[5]. These methods offer a dramatic increase in the rate of analysing null mutations compared with the traditional single gene approach. In addition, these novel approaches provide valuable information concerning gene expression[4], gene product localization[4] and the functional boundaries of ORFs[5]. However, using these techniques, thousands of ORFs must be individually examined following each selective growth condition to determine which mutants have displayed modified growth characteristics. This labour-intensive analysis step will make it difficult to test a large number of selective conditions on a genome-wide level.

Some of these difficulties have been overcome by a method in which transposons containing unique tags are used to generate insertional mutants[6]. The distinguishing tags allow collections of mutants to be pooled and analysed in parallel through selective growth assays. This approach has the advantage that all three steps (gene disruption, selection and analysis) are performed in parallel. Unfortunately, only 96 strains can be analysed in a single pool due to limitations of the filter-based hybridization assay[6]. In addition, the random nature of the transposon-based mutagenesis requires cloning and DNA sequencing to identify ORFs that are disrupted in strains with interesting phenotypes.

Here, we describe two novel approaches termed "quantitative phenotypic analysis" and "molecular bar-coding," in which directed gene replacement is used to generate individual deletion strains that are each labelled with a distinguishing molecular tag (unique 20 base-pair sequence). Tagged strains are pooled and analysed in parallel through selective growth conditions. The level at which each strain survives a given competitive growth condition can be determined by hybridizing the tags to a high-density oligonucleotide array. This approach has

[1]*Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA*
[2]*Affymetrix, 3380 Central Expressway, Santa Clara, California 95051, USA*

Correspondence should be addressed to R.W.D.

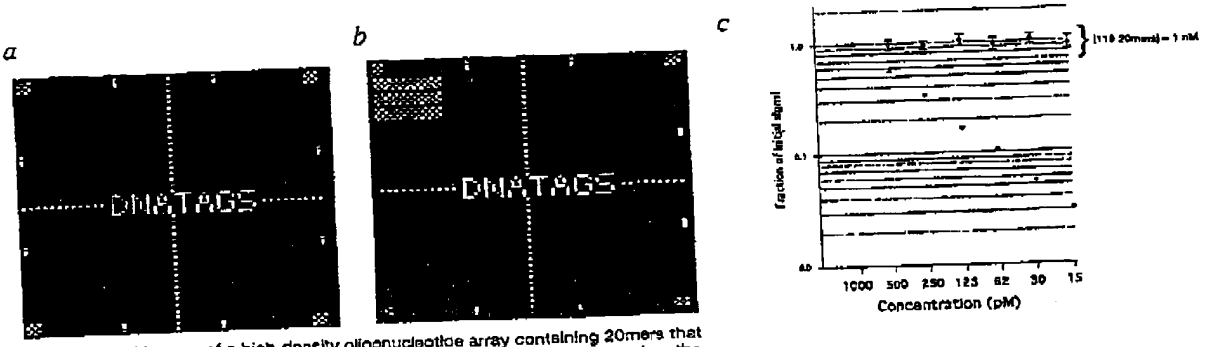


Fig. 1 Scanned images of a high-density oligonucleotide array containing 20mers that are complementary to 4,500 tag sequences. *a*, Hybridization pattern obtained when the array is hybridized with the fluorescently labelled 20mer that is complementary to the control oligonucleotide. Dark areas correspond to the synthesis sites for the 4,500 molecular tags. *b*, Hybridization pattern obtained when the array was hybridized with 120 different fluorescently labelled 20mers along with the control oligonucleotide. *c*, Plot showing the normalized data from a titration study using the pool of 120 fluorescently labelled oligonucleotides. Seven hybridization experiments were performed in which the concentration of one oligonucleotide was varied (1 nM–15 pM) while the remaining 119 oligonucleotides were held constant at 1 nM.

four main advantages: i) large pools of tagged deletion strains (potentially >6,000) can be simultaneously analysed because of the increased sensitivity afforded by fluorescent detection on high-density arrays; ii) deletion strains have ORFs completely deleted, avoiding residual or altered functions associated with truncated products; iii) deletion strains with interesting phenotypes are directly identified without need for further cloning or sequencing; and iv) the quantitative nature of the chip hybridization assay should make it possible to reveal subtle phenotypes that may otherwise be missed.

The steps of molecular bar-coding are as follows: i) individual deletion strains are generated by replacing specific ORFs with a selectable marker that is linked to a 20 base-pair molecular tag that serves as a unique identifier; ii) large numbers of tagged deletion strains are pooled and grown competitively under various selective conditions (for example, minimal medium, exposure to X-rays, high salt, and so on); iii) the molecular tags are amplified from the surviving strains using a common set of primers and hybridized to a high-density array containing, at defined positions, known oligonucleotides that are complementary to the tag sequences; and iv) the intensities of the hybridization signals for the tags on the array are measured and used to determine the relative abundance of the corresponding deletion strains in the pool. Measuring the relative abundance of the tags at different times during selective growth allows the fitness of each deletion strain in the pool to be quantitatively determined.

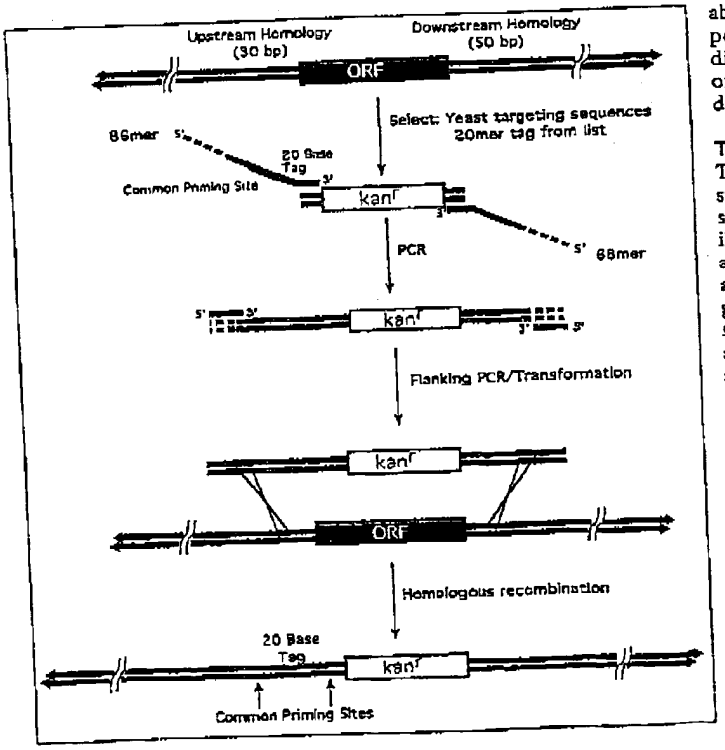


Fig. 2 PCR-targeting strategy used to generate tagged deletion strains. A dominant selectable marker (*kan^r*) was amplified using a pair of long primers that contained yeast sequences on the 5' end and homology to the marker on the 3' end. One of the oligonucleotides is a 68mer that contains 50 bases of yeast homology and 18 bases of homology to the marker. The other oligonucleotide is an 86mer that contains a 20 base tag and an 18 base tag priming site in addition to the 30 bases of yeast homology and the 18 bases of marker homology. The dashed lines represent missing sequences on the 5' ends of the long unpurified oligonucleotides (see Methods). A second round of PCR is performed with 20mers that are homologous to the ends of the initial PCR product to increase the amount of full length product. The product from the second round PCR is transformed into a haploid yeast strain and homologous recombination results in the replacement of the targeted ORF with the marker, 20mer tag, and tag priming site.

Tag design

The molecular tags are 20 base-pair DNA sequences specifically designed to serve as unique identifiers. Tag sequences are as different as possible yet still retain similar hybridization properties to facilitate simultaneous analysis on high-density oligonucleotide arrays. We used an algorithm to select a set of 9,105 maximally distinguished 20mer tag sequences that are predicted to have similar melting temperatures (61 ± 5 °C), no secondary structure and no extensive similarity between any two sequences in the list (>5 mismatches). This set of tags is sufficient for the entire yeast genome.

High-density oligonucleotide arrays

Hybridization to a high-density oligonucleotide array can detect the presence of a fluorescently labelled tag. We used a combination of photolithography and oligonucleotide synthesis chemistry to generate arrays containing thousands of oligonucleotides complementary to the different 20mer tag sequences[7,8]. Each element on the array measures 130 × 130 μm and contains approximately 10^8 oligonucleotides of a given sequence[9]. The location and identity of each sequence on the two-dimensional glass surface is known. Fluorescently labelled DNA is applied to the array and scanning confocal microscopy detects specific hybridization events. A single experiment (30 minute hybridization + 15 minute scan) generates quantitative hybridization information for thousands of different sequences.

To determine the optimal length for the tag sequences, we synthesized and tested high-density arrays containing thousands of specifically designed 8mers, 15mers, and 20mers (1,025, 8,000, and 4,500 oligonucleotides, respectively). The specificity and intensity of the hybridization signals obtained with 20mers were consistently superior to those obtained with the shorter oligonucleotides (data not shown).

For this study, we synthesized high-density arrays containing 20mers complementary to 4,500 of the specially designed tag sequences and synthesized control sequences in the corners, in a cross-hair pattern across the array and in a 'DNA TAGS' configuration. Fig. 1*a* shows a scanned image of an array hybridized with a fluorescently labelled 20mer complementary to the control sequences on the array. The dark areas on the image correspond to the synthesis sites for the 4,500 tags that did not cross-hybridize with the control oligonucleotide.

To test the hybridization properties of the tag sequences, we synthesized 120 fluorescently labelled oligonucleotides that are complementary to specific tag sequences on the array. These fluorescently labelled 20mers should generate a 10 × 12 checker-board hybridization pattern in the upper left-hand corner of the array. The predicted hybridization pattern was obtained with virtually no cross-hybridization outside the 120 targeted array elements (Fig. 1*b*). In this hybridization experiment, the signal intensity for some of the tags is higher than others (Fig. 1*b*). These highly reproducible differences reflect either the sequence dependence of the hybridization reaction and/or variation in the quality of the 120 synthetic, labelled oligonucleotides.

To determine whether the hybridization signals on the array accurately reflect the concentrations of tags in a population, we performed a titration experiment using the pool of 120 labelled oligonucleotides. The concentration of one oligonucleotide was varied in a series of seven hybridization experiments (1 nM, 0.5 nM, 0.25 nM, 0.12 nM, 60 pM, 30 pM, and 15 pM), while the concentrations of the remaining 119 oligonucleotides were held constant at 1 nM. A two-fold change in concentration was detected in the presence of this complex hybridization mixture (Fig. 1*c*). In addition, the hybridization intensities for the titrated oligonucleotide decreased in a linear fashion over the 70-fold concentration range tested (1 nM to 15 pM).

Table 1 Transformation results for auxotrophic ORFs

ORF name	ORF size	Targeting efficiency
ADE1 (M61209)	305aa	100% (8/8)
ADE2 (M59824)	571aa	50% (4/8)
ADE3 (M24737)	945aa	50% (4/8)
ADE4 (M74309)	509aa	100% (8/8)
ADE5 (X04337)	802aa	100% (8/8)
ARO2 (X60190)	376aa	75% (5/8)
ARO7 (M24517)	258aa	100% (8/8)
TRP2 (K01386)	507aa	100% (8/8)
TRP3 (K01386)	484aa	88% (7/8)
TRP4 (X04273)	380aa	75% (8/8)
TRP5 (V01342)	707aa	88% (7/8)

The ORF name (GenBank accession number), ORF size and targeting efficiency (correct integrations versus total) are shown for each of the 11 auxotrophic ORFs.

Generating tagged deletion strains

We used a variation of a previously described PCR targeting strategy[3,10] to generate the tagged deletion strains (Fig. 2). Normally, a selectable marker (such as *kan*^{r}) is amplified using a pair of long primers that contain yeast targeting sequences on the 5' end and homology to the marker on the 3' end. In our modified scheme, a 20-base tag and an 18 base tag priming site are incorporated into one of the oligonucleotides. The PCR-amplified selectable marker is transformed into a haploid yeast strain and homologous recombination results in the replacement of the targeted ORF (start codon to stop codon) with the marker, 20mer tag, and tag priming site. The integrated tag serves as a unique, permanent identifier for that deletion strain.

Pilot study

Gene disruption. To test the PCR targeting strategy (described in Fig. 2), we replaced 11 auxotrophic genes with deletion cassettes containing different pre-selected molecular tags. For each of the transformations, we analysed several G418 resistant colonies using a PCR assay that differentiates between correct and incorrect integration events[11]. Table 1 shows the transformation results for the 11 auxotrophic genes. The targeting efficiency ranged from 50–100% and there was no correlation between the correct integration frequency and the size of the targeted ORF.

Competitive growth. We tested the feasibility of using the molecular tags to monitor a population of cells using the auxotrophic deletion strains described in Table 1. A deletion pool was generated by combining equal numbers of cells from each of the 11 tagged deletion strains. Genomic DNA extracted from yeast cells in the deletion pool served as the template for a tag amplification reaction. This amplification step serves to increase the concentration of the tags and reduce the complexity of the sample hybridized to the array (mixture of amplified tags versus total yeast genomic DNA). The amplified tags were hybridized to the high-density array which was then washed and scanned. The signals for the 11 tags were intense and highly specific (Fig. 3). Only one out of the 131 tags tested so far has displayed a significant level of cross-hybridization (Fig. 3, yellow asterisk).

The 11 auxotrophic deletion strains provide a good system for testing the bar-coding strategy because the phenotypes are known. In a competitive growth experiment, the relative abundance of these strains should be maintained when the deletion pool is cultivated in complete medium because the strains are predicted to have

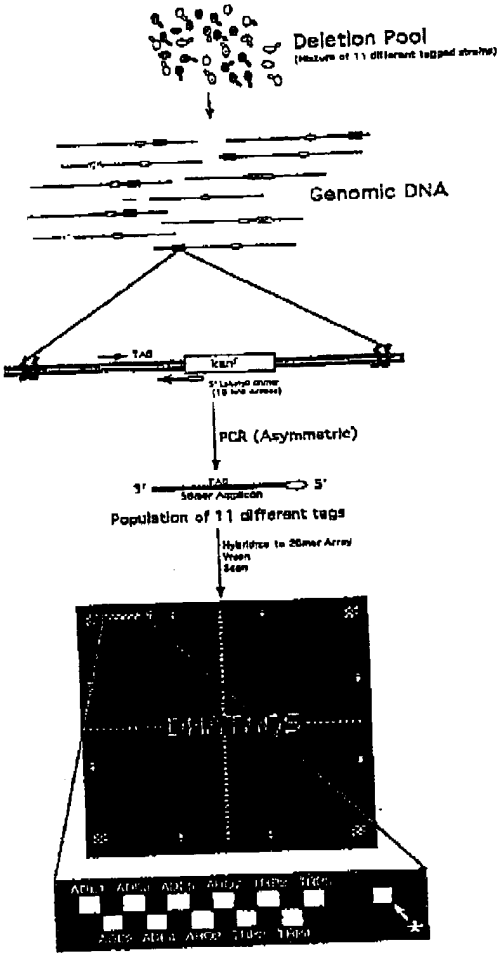


Fig. 3 Tag amplification strategy. Genomic DNA isolated from a deletion pool containing 11 tagged deletion strains was used as template for the tag amplification reaction. The tags were amplified using a pair of primers that were complementary to the common priming sites. The asymmetric PCR generates a mixture of single-stranded fluorescently labelled 56mer tag amplicons that were directly hybridized to the high-density array. A scanned image of the array is shown. A close-up view of the left hand corner shows the exact location of the 11 tags on the array. The asterisk indicates a cross-hybridizing sequence (see text).

similar growth rates. On the other hand, the *ade1*Δ–*ade5*Δ deletion strains or the *trp2*Δ–*trp5*Δ and *aro2*Δ deletion strains should be depleted from the population when the pool is grown in media lacking adenine or tryptophan, respectively. To test this prediction, we cultivated representative aliquots of the deletion pool in complete synthetic medium (SDC), medium lacking adenine (SDC-ADE), or medium lacking tryptophan (SDC-TRP). At various time points after inoculation, tags were isolated from the pool and analysed on the array to determine the relative abundance of the different deletion strains in the pool. Similar hybridization patterns were obtained at each of the time points when tags were isolated from the deletion pool grown in complete medium (Fig. 4*a*). In contrast, the signal intensities of the tags associated with the *ade1*Δ, *ade2*Δ, *ade3*Δ, *ade4*Δ, and *ade5*Δ deletion strains became progressively weaker when the pool was grown in medium lacking adenine (Fig. 4*b*). The expected depletion of the five Trp$^-$ deletion strains (*trp2*Δ, *trp3*Δ, *trp4*Δ, *trp5*Δ and *aro2*Δ) occurred when the pool was grown in medium lacking tryptophan (Fig. 4*c*). Figure 4*d* shows a quantitative representation of the hybridization data from the growth study in medium missing adenine. We measured the signal intensities of the *ade1*Δ and *trp5*Δ deletion strains at various time points during the growth study and plotted the normalized hybridization data (Fig. 4*d*). The graph clearly shows that the relative abundance of the *trp5*Δ deletion strain was maintained through the competitive growth while the *ade1*Δ deletion strain was depleted from the population (see Methods: data analysis).

Colony colour assay

We used an independent approach to determine whether the signal intensities of the tags on the array accurately reflect the relative abundance of different deletion strains in the pool. The *ade1*Δ and *ade2*Δ deletion strains form red colonies when plated on medium containing limiting amounts of adenine while the other nine deletion strains form white colonies on this medium. Determining the percentage of red colonies provides a direct measurement of the relative abundance of the *ade1*Δ and *ade2*Δ deletion strains in the pool. We cultivated representative aliquots of the deletion pool in complete medium and medium lacking adenine, and at each of the indicated time points, plated several thousand cells from the deletion pool. As predicted, 18% of the colonies taken from the starting population (2/11 deletion strains) were red. This percentage of red colonies remained constant when the deletion pool was grown in complete medium (SDC) for 10 population doublings. In contrast, the percentage of red colonies decreased when the deletion pool was grown in medium lacking adenine (Fig. 4*e*). The similarity between the plots shown in Figs 4*d* and 4*e* verifies that the hybridization signals on the array reflect the actual abundance of the tagged deletion strains in the pool.

Discussion

Determining the biological function for many of the newly identified yeast ORFs will require that thousands of different deletion strains be tested under a large variety of selection conditions[1]. Molecular bar-coding is ideal for this task because it allows large numbers of tagged deletion strains to be analysed simultaneously in a highly quantitative fashion. The labour-intensive (genome-wide) step of making the thousands of deletion strains will only have to be performed once, after which all subsequent selections and analyses could be performed in parallel. Furthermore, the quantitative nature of this approach may make it possible to observe subtle growth differences that might otherwise be overlooked. For example, the relative abundance of a single deletion strain with a 5% growth defect (doubling time of 147 minutes versus 140 minutes), grown in a competitive environment with 5,999 other deletion strains

article

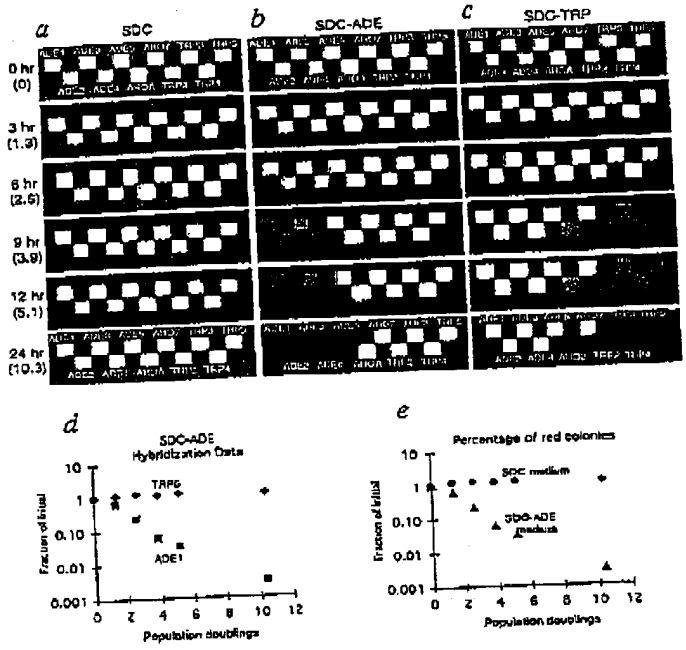


Fig. 4 Competitive growth study using a pool containing 11 tagged auxotrophic deletion strains. *a*, A representative aliquot of the deletion pool was grown in complete synthetic medium (SDC) and tags were isolated at the indicated times (population doublings are shown in parentheses) and analysed on the array. The upper-left hand corner of the array is shown for each of the time points *b*, A similar growth study performed in medium lacking adenine (SDC-ADE) and *c*, medium lacking tryptophan (SDC-TRP). *d*, Plot showing the normalized values for the *ade1Δ* and *trp5Δ* deletion strains from the growth study in medium lacking adenine. *e*, A plot showing the data from the colony colour assay. Aliquots of the deletion pool were grown in complete medium (SDC) and medium missing adenine (SDC-ADE). The relative abundance the *ade1Δ* and *ade2Δ* deletion strains in the pool was determined at each of the indicated time points by measuring the percentage of red colonies. The percentages from each time point were normalized to the values observed in the starting pool.

with wild-type growth rates (doubling times of 140 minutes), would be reduced by 30% after 10 population doublings and by over 85% after 60 population doublings. This dramatic change in the relative abundance should be easy to detect using the chip-based hybridization assay. Finally, the collection of individual deletion strains generated by this project would be immediately available to: i) confirm phenotypes observed in the pool; ii) make double mutants; and iii) cross into different genetic backgrounds.

Expanding this approach to the genome-wide level will require a high-throughput strategy to generate the thousands of tagged deletion strains. Towards this goal, a robust PCR deletion strategy has been developed in which all of the steps are either automated or performed in 96-well plates. A 96-well automated multiplex oligonucleotide synthesizer (A.M.O.S.) has also been developed to make the thousands of required oligonucleotides[12]. This machine has been designed to minimize the labour and reagent consumption involved in oligonucleotide synthesis. Over 20,000 oligonucleotides have been synthesized on a single machine in one year at a cost of 12 cents per base. The 42,000 oligonucleotides that will be required to delete all of the ORFs in the yeast genome (one 86mer, one 68mer and five 18mers for each of the 6,000 yeast ORFs) could be synthesized in less than a year using the three 96-well synthesizers currently in operation in our labouratory.

Developing a system capable of analysing thousands of tagged deletion strains is also essential to scaling this approach to the genome-wide level. The factors that must be addressed are tag amplification, array density, sensitivity of the detection system and cross-hybridization. The problem of biasing the relative levels of the different tags during the amplification step has been minimized by using a common pair of primers and by the fact that all of the tags are the same length, have similar base composition and are free of secondary structure. The existing light-directed synthesis technology can generate high-density arrays containing over 65,000 different oligonucleotides[13], which is well beyond the requirements of this project. Furthermore, the current fluorescent detection system can readily measure concentrations below 1 pM, which should be sufficient for monitoring pools containing more than 6,000 deletion strains (see Methods; pool size). Finally, the initial hybridization results with the 131 tags suggest that cross-hybridization is minimal and largely avoidable. All of the tags will be tested for cross-hybridization before use in the deletion strategy.

The goal of this project is to provide experimental data for thousands of deletion strains in a rapid and cost effective manner. These data, along with the collection of individual deletion strains, will be a valuable resource for investigating the function of genes in yeast.

Additional applications of the molecular bar-coding strategy include the generation of a complete set of tagged deletions in a 'true' wild-type yeast strain, one that has been recently isolated from a known ecological niche. The collection of deletion strains could be released back into their natural environment and the population monitored over time by surveying the PCR-amplifiable tags. This type of analysis would be useful for determining the activity of genes as a function of the environment. Molecular bar-coding may become an invaluable tool in this emerging field of 'molecular ecology'.

The molecular tags will also facilitate the task of keeping track of the thousands of deletion strains. Normally, different deletion strains are identified by labels (or barcodes) that are placed on the outside of the tubes or micro-titer plates. Incorrect labelling of tubes or putting the wrong strain in the wrong tube are serious problems that can be minimized by having a unique identifier in the actual deletion strains themselves. The identity of any strain could be confirmed by performing a simple tag-specific PCR or by analysing the tag on an array. Furthermore, the fact that the tags are physically linked to the mutations in the deletion strains may make it possible to perform recombination experiments on a genome-wide level. The linked nature of the molecular tags may also allow the entire collection of deletion strains to be mated into different genetic backgrounds. However, many applications only require that the molecular tag be in the same cell as the mutation. For example, tags could be introduced on a plasmid or into a

common region of the genome to generate a population of tag-containing wild-type cells. This mixture could be randomly mutagenized and strains with useful phenotypes (such as temperature sensitive mutations) could be identified using the molecular tags. Alternatively, tags could be individually introduced into an existing collection of mutants.

The molecular bar-coding strategy is not limited to analysing deletion strains in yeast. This approach can be applied to any task, *in vitro* or *in vivo*, that requires large populations (cells, DNA fragments, molecules, and so on) to be monitored in parallel. The tags can be introduced in a directed fashion as described, or randomly using either DNA cloning techniques or transposons. Approaches that allow parallel analysis of complex populations will be essential tools as researchers begin to study organisms on the genome-wide level.

Methods

Selection strategy to identify the tag sequences. Starting with all possible 20mers (1.2×10^{12}), a computer algorithm was used to eliminate sequences with known hybridization problems (for example, secondary structure, runs of single nucleotides, unbalanced base composition, and so on). Sequences containing common 9mers (or more) were eliminated during this step because contiguous stretches of homology (≥10 bp) have been shown to cause cross-hybridization (data not shown). The output of this programme was a list of 51,082 20mers that are predicted to be as different as possible yet still have similar melting temperatures. For increased specificity, the resulting set of 51,082 20mers was passed through a second filter in which a pair-wise analysis strategy was used to identify and eliminate the sequences most likely to cross-hybridize. Performing the pair-wise analysis using higher thresholds generated sets with fewer but more specific 20mers. Sets containing 51,082, 9,105, 2,643, 853, 170 and 42 20mers were obtained as the stringency level was incrementally increased. The 4,500 20mers described here were selected from the set containing 9,105 20mers.

Hybridization to the high-density oligonucleotide array. Fluorescently labelled oligonucleotides, complementary to 120 different tag sequences, were synthesized using an A.M.O.S. 96-well oligonucleotide synthesizer[12]. Yields were quantified by measuring absorbance at 260 nm and the oligonucleotides were used without purification in hybridization assays. The 200-μl hybridization mixture (6x SSPE-T) contained: 0.9 M NaCl, 60 mM NaH_2PO_4, 6mM EDTA and 0.005% Triton X-100. In addition, the solutions contained 0.5 nM control oligonucleotide (5'–fluorescein-CTGAACGGTAGCATCTTGAC–3'), and 1 nM of each of the 120 fluorescently labelled 20mers. Following a 20 min hybridization at 37 °C, the array was washed 10 times with 1× SSPE-T at 22 °C. The wash solution was removed and the hybridization chamber was re-filled with 6x SSPE-T prior to scanning. Hybridization events were detected by using an argon ion laser to excite the fluorescein; the resulting emission was detected using a photomultiplier tube through a 530 nm bandpass filter (Molecular Dynamics). The entire array was read at a resolution of 11.25 μm in less than 15 min, generating quantitative signal for each of the 4,500 tag sequences. The collected data was analysed using GeneChip software (Affymetrix).

Construction of the pFA-TAG plasmid. A PCR cloning strategy was used to introduce an 18-bp tag priming site into the multiple cloning region of the pFA6-kanMX2 module[10]. A 1.3-kb PCR product was generated using the pFA6-kanMX2 module as template and a pair of primers complementary to the regions flanking the marker (underlined sequences). The tag priming site is included as an 18-base extension on the 5' end of one of the primers (5'–CTATAGTGAGTCGTATTAAAG-CTTCGTACGCTGCAG–3'). The other primer does not have any additional sequence information on the 5' end (5'–GGG-CACTAGTGGATCTGA–3'). The reaction mixture contained: 20 mM Tris (pH 8.8), 10 mM KCl, 10 mM $(NH_4)_2SO_4$, 2.0 mM $MgCl_2$, 0.1% Triton X-100, 250 μM each dNTP, 0.2 μM each primer and 2.5 U Vent DNA polymerase (New England Biolabs) in a 100 μl total volume. The conditions were as follows: 94 °C for 3 min, then 25 cycles of 94 °C for 15 s, 50 °C for 15 s, 72 °C for 30 s, followed by a 3 min extension at 72 °C. The gel-purified PCR product was cloned into *Sma*I linearized pBluescriptSK+II (Stratagene) and transformed into electrocompetent DH5α cells (GibcoBRL). White colonies were picked from an LB plate (100mg/l ampicillin, IPTG and X-gal) and analysed using PCR. A *kan*^{r} specific primer (5'–CCTCGACAT-CATCTGCCC–3') along with vector-specific primers (–20 and reverse) were used to confirm the presence and orientation of the insert. The resulting pFA-TAG plasmid was partially sequenced to determine if any mutations were introduced during the cloning procedure.

PCR-targeting strategy. The pFA-TAG module was amplified using a pair of primers that have 3' ends (18 bases) homologous to the marker and 5' ends (30–50 bases) that contain yeast targeting sequences[9,10]. One of the long oligonucleotides also contained a unique 20mer tag and 18mer common tag priming site (Fig. 2). The *kan*^{r} gene (aminoglycoside phosphotransferase) from the bacterial transposon *Tn*903 was used as the dominant selectable marker[10]. Expression of the *kan*^{r} gene confers resistance to the drug G418 in yeast[14]. The 100 μl PCR reaction contained: 2.5 U of *Taq* DNA polymerase, 10 mM Tris (pH 8.5), 50 mM KCl, 1.5 mM $MgCl_2$, 0.2 μM each primer, 250 μM each dNTP and 10 ng of pFA-TAG template. The reaction conditions were: 3 min at 94 °C, then 30 cycles of 15 s at 94 °C, 15 s at 50 °C and 30 s at 72 °C followed by a 3 min extension at 72 °C. Due to limits in oligonucleotide synthesis chemistry, the proportion of full-length 86mers in an unpurified sample is low. PCR products generated with primers of this length result in low transformation efficiencies because of the missing yeast targeting information on the ends (Fig. 2; blue dashed lines). Purification by HPLC, although suitable, is labour-intensive and impractical to perform on large numbers of oligonucleotides. Instead, the amount of full-length PCR products was increased through a second round of PCR using 20mers that are homologous to the ends of the first PCR product. The second round flanking primer PCR was performed using 1 μl of the first PCR as template and a pair of 20-base primers that are homologous to the ends of the yeast targeting sequences (Fig.2). The amplification conditions for the flanking PCR are identical to the conditions used in the first round. The PCR product was analysed on a 1% agarose gel and used without purification in the subsequent transformation.

Transformations. Yeast growth medium and the standard techniques for manipulating yeast have been described[15]. The haploid yeast strain YJM826 [*MAT*α *SUC2 CUP1* Gal$^+$], isogenic with S288C, was used for the transformations (provided by John McCusker). Cells (1×10^8) were transformed with approximately 1 ug of a linear PCR product using a standard lithium acetate method[16]. Transformed cells were grown at 30 °C in YPD for 4 h and plated on YPD plates containing 300 mg/l G418 (Gibco BRL). Large G418-resistant colonies (5 to 50) appeared after 3 d growth at 30 °C and the background was low. Colonies (8) from each transformation were streaked onto fresh plates and single colonies were isolated and analysed using PCR with three primers[11]. In addition, the putative colonies were checked by replica plating onto the appropriate drop-out medium for each of the 11 different auxotrophic ORFs (data not shown). The confirmed deletion strains were stored in 15% glycerol at –70 °C. The 11 auxotrophic deletion strains described in Table 1 were all complete deletions (start codon to stop codon).

Generating the deletion pool. Separate 20-ml liquid YPD cultures were inoculated with each of the 11 different auxotrophic tagged deletion strains. The cultures were grown at 30 °C until mid-exponential phase (0.5–1.0 OD_{600}). An equal number of cells were combined to generate a master deletion pool which was stored in 10-OD_{600} aliquots in 15% glycerol at –70 °C.

Tag amplification and hybridization. The asymmetric tag amplification reaction was performed on 1 ug of genomic DNA isolated from a mid-exponential phase culture of the deletion pool as previously described[17]. One of the common tag amplification primers is 5' labelled with 5(6)-carboxyfluorescein and included in 10-fold excess over the unlabelled primer in this asymmetric PCR. The 100-µl reaction mixture contained: 10 mM Tris (pH 8.5), 50 mM KCl, 1.5 mM $MgCl_2$, 0.5 µM F-common 1 (5'–fluorescein-GCTTTAATACGACTCACTATAG–3'), 0.05 µM common 2 (5'–GATGTCCACGAGGTCTCT–3'), 250 µM each dNTP, and 2.5 U AmpliTaq DNA polymerase (Perkin-Elmer). The conditions were as follows: 94 °C for 3 min, then 30 cycles of 94 °C for 15 s, 50 °C for 15 s, 72 °C for 30 s, followed by a 3 min extension at 72 °C. The single-stranded, fluorescently labelled 56mer tag amplicons were directly hybridized to the array without purification. The 200-µl hybridization mixture contained: 6x SSPE-T, 0.5 nM control oligonucleotide and 50 µl of the tag amplification reaction (~25 nM mixture of 11 Tags; 2.2 nM each). The hybridization, washing and scanning conditions were as described above.

Deletion pool growth study. An aliquot of the deletion pool was cultivated in synthetic complete (SDC) medium at 30 °C and aliquots containing 3×10^7 cells were harvested from the culture at 0, 3, 6, 9, 12 and 23 h for subsequent analysis on the high-density array. The doubling time for the population is approximately 140 min in SDC medium. To ensure that the cells were harvested from mid-log phase cultures at each of the time points, a series of 2-fold dilutions of the deletion pool were made at the beginning of the growth experiment. Specifically, 10 ml cultures containing 3×10^8, 1.5×10^8, 7.5×10^7, 3.8×10^7, 1.9×10^7, 9.4×10^6, 4.7×10^6 and 2.3×10^6 cells were generated by making 2-fold serial dilutions with SDC medium. At each time point, ~3×10^7 cells were harvested from a culture that was closest to mid-log phase (between 0.5–1.0 OD/ml). This procedure was repeated using SDC-ADE and SDC-TRP medium.

Data analysis. The growth rates of the tagged deletion strains in the pool are measured relative to a reference strain. The reference strain has a tag integrated into a region of the genome that is predicted not to affect the fitness of the cell. At each time point, the signal intensities for each of the elements on the array were measured. After subtracting the background, the values were normalized to the signal from the reference strain to account for variations in the tag amplification and hybridization conditions. The 'fraction of initial' values were obtained by dividing the normalized values from each time point by the normalized values from the initial time point. A ratio of one indicates that the relative abundance of that deletion strain has not changed in the population. A ratio less than one indicates that the deletion strain was depleted from the population.

In the plot shown in figure 4d (ADE minus medium), the *trp2Δ* deletion strain was used as the reference strain. Similar results were obtained when any of the other non-ADE deletion strains (e.g. *aro2Δ*, *aro7Δ*, *trp3Δ*, *trp4Δ*, and *trp5Δ*) were used as the reference strain. In the future, several different control strains will be constructed and included in the deletion pool. These strains will have tags integrated into regions of the genome that should not affect the fitness of the cell (in pseudo genes, duplicated genes, intergenic regions and so on).

Colony colour assay. An aliquot of the deletion pool was diluted as described above and grown in synthetic complete medium. At various time points (0, 3, 6, 9, 12 and 23 h) several thousands of cells were taken from the pool and plated on rich medium that was not supplemented with adenine. The number of red colonies (*ade1Δ* and *ade2Δ* deletion strains) and white colonies (other nine deletion strains) was determined for each of the time points. The percentage of red colonies was determined at each time point and the values were normalized to the percentage of red colonies that were observed in the starting pool (such as 18%). This process was repeated in medium missing adenine.

Pool size. The concentration of a tag in a hybridization mixture is inversely proportional to the total number of tags in the pool. Tags amplified from a pool containing 6,000 strains would result in an individual tag concentration of 85 pM (100 pmoles can be generated by a single tag amplification reaction/ 6,000 strains = 17 fmoles of each tag which is hybridized in a volume of 200 µl). This concentration is well within the detection limits of the fluorescent scanner.

Acknowledgements

We thank T. Gingeras, S. Fodor, E. Hubble, E. Gentalen, D. Lockhart, R. Sapolski, and M. Chee from Affymetrix; E. Winzeler, W. Shih, M. Schena, R. Cho, M. Campbell, J. McCusker, D. Richards, and G. Cavet for helpful discussions and a critical reading of the manuscript.

Received 23 August; accepted 1 November 1996.

1. Oliver, S.G. From DNA sequence to biological function. *Nature* 379, 597–600 (1996).
2. Rothstein, R. Targeting, disruption, replacement, and allele rescue: integrative DNA transformation in yeast. *Meth. Enzymol.* 194, 281–301 (1991).
3. Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. & Cullin, C. A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae. *Nucl. Acids Res.* 21, 3329–3330 (1993).
4. Burns, N. *et al.* Large-scale analysis of gene expression, protein localization, and gene disruption in Saccharomyces cerevisiae. *Genes Dev.* 8, 1087–1105 (1994).
5. Smith, V., Botstein, D. & Brown, P.O. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc. Natl. Acad. Sci. USA* 92, 6479–6483 (1995).
6. Hensel, M. *et al.* Simultaneous identification of bacterial virulence genes by negative selection. *Science* 269, 400–403 (1995).
7. Fodor, S.P. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767–773 (1991).
8. Pease, A.C. *et al.* Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* 91, 5022–5026 (1994).
9. Lockhart, D., J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotech.* (in the press).
10. Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. *Yeast* 10, 1793–1808 (1994).
11. Huxley, C., Green, E.D. & Dunham, I. Rapid assessment of S. cerevisiae mating type by PCR. *Trends Genet.* 6, 236 (1990).
12. Leamkerl, D.A., Hunicke-Smith, S.P., Norgren, R.M., Davis, R.W. & Brennan, T. An automated multiplex oligonucleotide synthesizer: development of high-throughput, low-cost DNA synthesis. *Proc. Natl. Acad. Sci. USA* 92, 7912–7915 (1995).
13. Lipshutz, R.J. *et al.* Using oligonucleotide probe arrays to access genetic diversity. *BioTech.* 19, 442–447 (1995).
14. Jiminez, A. & Davies, J. Expression of transposable antibiotic resistance elements in Saccharomyces *Nature* 287, 869–871 (1980).
15. Sherman, F., Fink, G.R. Hicks, J.B. *Methods in yeast genetics: A laboratory manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1986).
16. Gietz, R.D. & Woods, R.A. High efficiency transformation with lithium acetate in *Molecular Genetics of Yeast, A Practical Approach* (ed. Johnston, J.R.) 121–134 (IRL Press, Oxford, UK, 1994).
17. Hoffman, C.S. & Winston, F. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. *Gene* 57, 267–272 (1987).